Advances in computational ion mobility mass spectrometry; with application to α_1-antitrypsin

Ganesh Nadaraj Sivalingam

Institute of Structural and Molecular Biology
University College London

Thesis submitted for the degree of Doctor of Philosophy

September 2014

Declaration

I, Ganesh Nadaraj Sivalingam declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

A new method for deconvolution of electrospray ionisation mass spectrometry (ESI-MS) spectra was produced, allowing for the masses of overlapping charge state series to be correctly indentified. The algorithm also determines the abundance of individual molecular species with a much higher accuracy for congested spectra.

Several new methods for representing TWIM-MS data were developed. The combination of the deconvolution algorithm with travelling wave ion mobility data creates plots with collision cross section (CCS) axes which can be directly compared with X-ray crystallography structures and computational models. Difference plots have allowed multidimensional analysis of changes in condition, and spectral averaging can produce a single representative spectrum from multiple replicates.

Gas-phase unfolding experiments using TWIM-MS are a popular method for probing protein stability in response to conditions such as ligand binding. The algorithms for processing these data are however in their infancy. This thesis describes the first deconvolution algorithm for gas-phase unfolding data, allowing for the accurate interpretation of conformation cross sections and abundances during the unfolding procedure.

The methodologies developed were then applied to α_1-antitrypsin, a metastable, aggregation prone protein. The protein was bound to a ligand, Ac-TTAI-NH_2, which has been shown to block aggregation as a titration and the MS deconvolution method was used to quantify the abundances of each bound state in each mass spectrum.

The first use of IM-MS to analyse *ex vivo* aggregates are shown, and the ion mobility methods created were used to determine the CCS values of the monomeric and dimeric species. The interaction between α_1-antitrypsin and Ac-TTAI-NH_2 was probed using gas-phase unfolding experiments, determining that the ligand stabilises the protein, with a specific pattern of gas-phase unfolding observed for each state.

Acknowledgements

"There is a computer disease that anybody who works with computers knows about. It's a very serious disease and it interferes completely with the work. The trouble with computers is that you 'play' with them!"

Richard P. Feynman

Surely You're Joking Mr Feynman (1989)

I would first like to thank my supervisor Dr. Konstantinos Thalassinos, his tutelage has helped me in many ways, notably in regards to constructing the flow of scientific thought through reports and presentations. I would also like to thank my second supervisor Dr. Bibek Gooptu for setting up my collaboration with Sarah Faull of the Lomas group, which has been very fruitful.

I would also like to give a big thank you to Dr. Mark Williams, who was my MSc project supervisor, thesis chair and the most knowledgeable scientist I have ever met. The enormous amounts of time and effort he gave me during my Ph.D, even though I was not his student, was incredible and vital to my success.

I, of course, also need to thank Dr. Adrian Shepherd for teaching me how to write computer programs, as well as helping me to get the place on this Ph.D. programme. Also I would like to thank his student Skelton, who first told me about Ubuntu, the operating system that I've used ever since that day.

My time as part of the Thalassinos Lab has been made enjoyable by the people within it. I would like to thank Dr. Richard Kerr and Dr. Jun Yan for initially teaching me and Harpal how to use the instrument... which was not exactly easy. I also need to thank Jun for her continual guidance, she still helps me now even though she left the lab a year ago! Adam's almost infinite patience continues to amaze me, and I am very grateful for all the times he listened (and paid attention!) to me ranting about computer-related topics. Zoja Nagurnaja's views on the world have fascinated me and David

Sutherland's comments in lab meetings have always been insightful, so I would like to thank the two of them as well. Harpal Sahota has been a great friend throughout the highs and lows of our Ph.D.s and I'm going to miss our coffee breaks!

I would like to thank Sarah Faull, Dr. Imran Haq and Dr. James Irving from the Lomas group for providing me with α_1-antitrypsin samples. I owe a debt of gratitude to Dr. Peg Nyon for her help and patience with regards to my α_1-antitrypsin work.

There are several additional students I would like to thank. Irene Farabella has given me loads of great advice, including helping me decide to try mass spectrometry instead of synthetic chemistry... Flora Scott and Liz Rodriguez, for the good times and for entertaining computer-related questions. I also like to mention Tom Warelow for being a great guy and achieving the feat of doing his Ph.D., going drinking with us and looking after two kids.

Duncan Tait has been a great friend and we have had excellent scientific and computational discussions. It has been refreshing to have someone I from outside of university who understands what I do.

To my friends, especially Jo and Tom Lawson and Zoe Fullard, thank you for bearing with me during the write up period. I am happy to say, you will be seeing a lot more of me now.

I would like to give a huge thank you to my cousin, Kajendiran Mahendiran. He has helped me throughout my education, and has always made time to proof read my reports when asked.

Finally, I would like to thank my long-suffering parents Lalitha and Siva Sivalingam for supporting me through my many years of education.

Contents

1 A selective history of mass spectrometry **19**

2 Mass spectrometry **23**

2.1 Electrospray ionisation 23

2.1.1 Interpreting ESI mass spectra 26

2.1.2 Nanoelectrospray ionisation 27

2.1.3 Mass spectrometry peak shapes 29

2.1.4 Non-covalent mass spectrometry 31

2.1.5 Denaturing mass spectrometry 35

2.2 Mass analysis and ion detection 38

2.2.1 Linear quadrupole mass analyser 40

2.2.2 Time-of-flight mass analyser 41

2.2.3 Tandem mass spectrometry 45

2.2.4 Microchannel plate detector 48

3 Ion mobility mass spectrometry **49**

3.1 Drift tube ion mobility 50

3.2 Travelling wave ion mobility 52

3.3 Collision cross section calculations 60

3.4 Gas-phase protein conformation 63

3.5 TWIM-MS experiments 64

4 Amphitrite **75**

4.1 Introduction 75

4.1.1 Deconvolution of ESI mass spectra 75

4.1.2 Ion mobility mass spectrometry data analysis 78

4.2 Methods 80

4.2.1 Sample sources . . . 80
4.2.2 Sample preparation . . . 80
4.2.3 Capillary preparation . . . 81
4.2.4 nESI-MS calibration . . . 81
4.2.5 TWIM-MS . . . 82
4.2.6 Experimental procedures . . . 82
4.2.7 Software development . . . 83
4.3 Results and discussion . . . 85
4.3.1 Mass spectrum simulation . . . 85
4.3.2 ATD extraction . . . 88
4.3.3 Calibration . . . 89
4.3.4 Applying a calibration . . . 90
4.3.5 Complex mixture analysis . . . 92
4.3.6 Spectral averaging . . . 94
4.3.7 Comparing different conditions . . . 95
4.3.8 Collision induced unfolding . . . 96
4.3.9 Arsenite oxidase . . . 97
4.3.10 Current state of software development . . . 100
4.4 Conclusion . . . 101
4.5 Appendix . . . 108
4.5.1 Example maximum entropy (MaxEnt) spectrum . . . 108
4.5.2 Simplifying calibration equations . . . 109
4.5.3 Amphitrite graphical user interfaces . . . 111

5 Challenger 115
5.1 Introduction . . . 115
5.1.1 Gas-phase unfolding of proteins . . . 115
5.1.2 Aims . . . 120
5.2 Methods . . . 121
5.2.1 Sample sources . . . 121
5.2.2 Mass spectrometry sample preparation . . . 122
5.2.3 IM-MS procedures . . . 122
5.2.4 Genetic algorithms . . . 123
5.2.5 Software development . . . 124
5.3 Results and discussion . . . 125

5.3.1 Automation of unfolding experiments 125
5.3.2 Summarising IM-MS unfolding data 127
5.3.3 Challenger algorithm development 133
5.3.4 Challenger algorithm optimisation 138
5.3.5 Experimental data deconvolution 144
5.3.6 Algorithm limitation 146
5.3.7 Current software development state 149
5.4 Conclusion 150
5.5 Appendix 158

6 α_1-antitrypsin 161
6.1 Introduction 161
6.1.1 Pathology of severe α_1-antitrypsin deficiency 162
6.1.2 Naming of α_1-antitrypsin variants 163
6.1.3 Treatment of severe α_1-antitrypsin deficiency 164
6.1.4 Evolutionary benefit of α_1-antitrypsin deficiency 164
6.1.5 Loop-sheet polymerisation model 166
6.1.6 β hairpin polymerisation model 167
6.1.7 C-terminal domain swap polymerisation model 168
6.1.8 Peptides block polymerisation 169
6.1.9 Aims 170
6.2 Methods 171
6.2.1 Sample sources 171
6.2.2 Recombinant α_1-antitrypsin expression and purification 171
6.2.3 Ac-TTAI-NH_2 α_1-antitrypsin titration 172
6.2.4 Ac-TTAI-NH_2 α_1-antitrypsin unfolding experiments . . 172
6.2.5 M polymer preparation 173
6.2.6 Glycosylated ion mobility Z α_1-antitrypsin experiments 174
6.3 Results and discussion 176
6.3.1 Ac-TTAI-NH_2 titration with wild type α_1-antitrypsin . 176
6.3.2 Unfolding experiments 181
6.3.3 Analysis of *ex vivo* α_1-antitrypsin 186
6.3.4 Ion mobility analysis of *ex vivo* polymers 191
6.4 Conclusion 195
6.5 Appendix 205

6.5.1 Analysis of dissociated holo in CIU experiments 205
6.5.2 Ac-TTAI-NH_2 - α_1-antitrypsin titration 207
6.5.3 Tables of CCS values for *ex vivo* analysis 208

7 Conclusions **209**
7.1 Amphitrite 209
7.2 Challenger 210
7.3 α_1-antitrypsin 212
7.4 Final remarks 213

8 Publications **215**

List of Figures

1.1 Parabola spectrographs 20

2.1 Example ESI mass spectrum 24
2.2 Schematic diagrams of the electrospray ionisation process . . . 25
2.3 Comparison of mass spectrum peak shapes 29
2.4 Nanoelectrospray ionisation mass spectra of myoglobin in different sample buffers 32
2.5 Analysis of the interaction between GyrA59 and SD8 33
2.6 Monitoring the rate of subunit exhange of GlmS 35
2.7 ExsG at varying ammonium acetate concentrations 36
2.8 Mass analysis of unfolded AioB 37
2.9 Mass resolution and resolving power 39
2.10 Illustration of the workings of a quadrupole mass analyser . . 40
2.11 A 45 kDa protein analysed with different buffer conditions . . 44
2.12 Linear and reflectron ToF mass analysers 45
2.13 CID used to analyse a heterogeneous sample with heavily overlapped peaks 47

3.1 Schematic representation of ion mobility separation 49
3.2 Schematic representation of the Waters Synapt travelling wave ion mobility mass spectrometer 52
3.3 Schematic of travelling wave propagation, and the resulting arrival time separation 53
3.4 Example of the multidimensional data that can be acquired simultaneously using a Waters Synapt TWIM-MS instrument 55
3.5 Polarisability of IM buffer gases 56

3.6 A TWIM-MS calibration curve using denatured myoglobin as the calibrant 59
3.7 An illustration of the projection approximation (PA) algorithm 61
3.8 CIU fingerprints of avidin and ConA with $Mg(OAc)_2$ 67

4.1 Methods of data representation in ion mobility mass spectrometry 79
4.2 An m/z vs. arrival time plot showing the multidimensional data available from a TWIM-MS experiment 80
4.3 Mass spectrum calibration with MassDiff 82
4.4 Native nESI peak shape models 86
4.5 Different stages in extracting arrival time distribution plots of serum amyloid P (pentamer) using Amphitrite 87
4.6 Creation of a CCS calibration using Amphitrite 90
4.7 Charge state collision cross section plots of cytochrome c . . . 91
4.8 IM-MS analysis of a mixture of BSA, concanavalin A and alcohol dehydrogenase 93
4.9 Amphitrite spectral averaging demonstration 94
4.10 Subtraction plot comparing ADH at 20 °C and 60 °C 95
4.11 Amphitrite CIU analysis of cytochrome c 97
4.12 Deconvolution of an arsenite oxidase mass spectrum. 98
4.13 Diagram of the four potential assembly pathways of arsenite oxidase 98
4.14 Arsenite oxidase assembly pathway analysis 99
4.15 Maximum entropy spectrum produced by the MaxEnt 3 function of Waters MassLynx software 108
4.16 ImProcessorGui from Amphitrite 111
4.17 ApplyCalibrationGui from Amphitrite 111
4.18 AtroposGui from Amphitrite 112
4.19 CalibrationGui from Amphitrite 113
4.20 IesGui from Amphitrite 113
4.21 SpectralAveragingGui from Amphitrite 114

5.1 Gas-phase unfolding of apo and holo myoglobin 117
5.2 Gas-phase unfolding and conformational abundance tracking of wild type and L55P mutant tetrameric transthyretin 118

5.3 CIU fingerprint analysis 119
5.4 GUI for automatically generating mass spectrometer settings files (.ipr) for the Waters Synapt. 126
5.5 Arrival time distributions of the gas-phase unfolding of lysozyme, β-lactoglobulin and myoglobin. 127
5.6 Demonstration of summary statistic calculations 128
5.7 Unfolding curves for model proteins 129
5.8 Variability curves for model proteins 130
5.9 CIU fingerprints of model proteins 131
5.10 Graphical user interface for calculating and plotting summary statistics, as well as plotting CIU fingerprints 132
5.11 Comparison between execution times for the three implementations of the Challenger algorithm 136
5.12 Deconvolution algorithm run on synthetic collision energy ramp data 137
5.13 Optimising the mutation rate and crossover parameters, for experimental data 139
5.14 Optimising the population size for experimental lysozyme data 141
5.15 Deconvolution, abundance analysis and comparison to summary statistics of lysozyme gas-phase unfolding 144
5.16 Examples of an experimental ATD with poor fits 145
5.17 Demonstration of a limitation of the deconvolution algorithm when analysing data with many similar ATDs 147
5.18 Deconvolution of a reduced β-lactoglobulin dataset 148
5.19 Summary statistics GUI 149
5.20 Challenger algorithm deconvolution of myoglobin. 158
5.21 Deconvolution of β-lactoglobulin data when not fitting for conformational centres. 159
5.22 Using results of the Challenger algorithm to deconvolute lysozyme CIU fingerprints 160

6.1 X-ray crystallography structure of α_1-antitrypsin, with secondary structure elements labelled 162
6.2 The α_1-antitrypsin loop-sheet model of polymerisation 166
6.3 The α_1-antitrypsin β hairpin model of serpin polymerisation . 167

6.4 The α_1-antitrypsin C-terminal domain swap model 169
6.5 Proposed model for Ac-TTAI-NH_2 binding with α_1-antitrypsin 170
6.6 Amphitrite deconvolution of α_1-antitrypsin and Ac-TTAI-NH_2 mass spectrum 176
6.7 α_1-antitrypsin in complex with a protease 177
6.8 nESI-MS titration of Ac-TTAI-NH_2 and α_1-antitrypsin 178
6.9 Synthetic mass spectrum abundance analysis 179
6.10 Deconvoluted abundances of Ac-TTAI-NH_2, α_1-antitrypsin titration 180
6.11 Stacked CCSDs of apo and holo α_1-antitrypsin 181
6.12 CIU fingerprint analysis of apo and holo α_1-antitrypsin 182
6.13 Summary statistic analysis of apo and holo α_1-antitrypsin . . 183
6.14 Challenger results for apo and holo α_1-antitrypsin 185
6.15 Deconvolution of Z mutant α_1-antitrypsin mass spectrum . . . 187
6.16 Simulation of plasma Z α_1-antitrypsin with Ac-TTAI-NH_2 . . 188
6.17 Mass spectra of M α_1-antitrypsin after incubating at different polymerising conditions 189
6.18 Mass spectrum of Z mutant α_1-antitrypsin polymer extracted from human liver. 190
6.19 α_1-antitrypsin dimer models 191
6.20 Ion mobility extraction limits for Z mutant α_1-antitrypsin oligomers 193
6.21 Ion mobility analysis of *ex vivo* Z α_1-antitrypsin polymers . . . 194
6.22 Analysis of dissociated holo in the CIU experiment 205
6.23 Time course of wild type α_1-antitrypsin binding to Ac-TTAI-NH_2 207
6.24 Mass spectrum of Ac-TTAI-NH_2 bound to wild type α_1-antitrypsin 207

List of Tables

5.1 IM-MS settings table for unfolding experiments on model proteins. 123
5.2 Table showing changes in error over generations 142

6.1 Mass spectrometer settings used for; Ac-TTAI-NH_2 α_1-antitrypsin titration experiments, Z α_1-antitrypsin extracted from plasma, and M and Z oligomer experiments. 172
6.2 Ion mobility mass spectrometry settings used for Ac-TTAI-NH_2 α_1-antitrypsin gas-phase unfolding experiments and oligomeric Z α_1-antitrypsin extracted from hepatocytes. 173
6.3 Table showing the change in CCS from the most native-like analysis to a given collision energy as a percentage. 184
6.4 Table showing the weight standard deviation values of apo and holo α_1-antitrypsin at 3 collision energies. 184
6.5 Collision cross section values of α_1-antitrypsin conformations, as determined using the Challenger deconvolution algorithm. . 186
6.6 Table of collision cross section (CCS) values of monomeric and dimeric charge states of *ex vivo* polymer Z mutant α_1-antitrypsin, given as peak top values and weighted mean CCS values. 208
6.7 Table of the CCS values and percentage increase in CCS for three polymerisation models and experimental data of *ex vivo* Z mutant α_1-antitrypsin polymer. 208

List of Abbreviations

2D . . . . . Two-dimensional
Å . . . . . Ångstöm
ADH . . . . . Alcohol dehydrogenase
ASA . . . . . Accessible surface area
ATD . . . . . Arrival time distribution
BSA . . . . . Bovine serum albumin
C-terminal . Carboxyl-terminal
CA . . . . . California
CCS . . . . . Rotationally averaged collision cross section
CCSD . . . . Collision cross section distribution
cDNA . . . . Complementary deoxyribonucleic acid
CHAMP . . . Calculating heterogeneous assembly and mass spectra of proteins
CID . . . . . Collision induced dissociation
CIU . . . . . Collision induced unfolding
COPD . . . . Chronic obstructive pulmonary disease
CRM . . . . . Charged residue model
DC . . . . . Direct current
DT . . . . . Drift tube
DTIM . . . . Drift tube ion mobility
Da . . . . . Dalton
E. coli . . . . *Escherichia coli*
EDTA . . . . Ethylenediaminetetraacetic acid
e.g. . . . . . Exempli gratia (for example)
EHSS . . . . . Exact hard sphere scattering
ESI . . . . . Electrospray ionisation
et al. . . . . . *et alii* (and others)
FDA . . . . . United States Food and Drug Administration
FWHM . . . Full width half maximum
GB . . . . . Gigabyte
GNU . . . . . GNU's Not Unix!

GPGPU . . . General-purpose computing on graphics processing units
GPU . . . . . Graphical user interface
GuHCL . . . Guanidinium chloride
i.e. . . . . . Id est (that is)
IEM . . . . . Ion evaporation model
IM . . . . . Ion mobility
IPTG . . . . . Isopropyl β-D-1-thiogalactopyranoside
MALDI . . . Matrix assisted laser desorption ionisation
MCP . . . . . Microchannel plate
MO . . . . . Missouri
MS . . . . . Mass spectrometry
m/z . . . . . Mass-to-charge ratio
nESI . . . . . Nano-electrospray ionisation
NMR . . . . . Nuclear magnetic resonance
NY . . . . . New York
OD . . . . . Optical density
OS . . . . . Operating system
PA . . . . . Projection approximation
PAGE . . . . Polyacrylamide gel electrophoresis
PDB . . . . . Protein Databank
PSA . . . . . Projection superposition approximation algorithm
Pi . . . . . Protease inhibitor
ppm . . . . . Parts per million
Q-ToF . . . . Quadrupole time-of-flight
RAM . . . . . Random-access memory
RCL . . . . . Reactive centre loop
RF . . . . . Radio frequency
RNA . . . . . Ribonucleic acid
SAP . . . . . Serum amyloid P component
SLD . . . . . Soft laser desorption
SOMMS . . . Solving complex macromolecular mass spectra
SRIG . . . . . Stacked ring ion guide
TM . . . . . Trajectory method

TRAP . . . . *trp* RNA binding attenuation protein

TTR . . . . . Tetrameric transthyretin

TWIM . . . . Travelling wave ion mobility

ToF . . . . . Time-of-flight

UK . . . . . United Kingdom

USA . . . . . United States of America

USSR . . . . . Union of Soviet Socialist Republics

XML . . . . . Extensible markup language

Chapter 1

A selective history of mass spectrometry

Mass spectrometry is a technique that allows the accurate determination of the mass-to-charge ratio (m/z) of gas-phase ions, with the change in mass of 1 Da (the mass of a proton), in a 18,000 Da protein being detectable[1]. It is used in a wide variety of applications, ranging from the identification of illict compounds in the blood of olympic athletes [2] to analysing intact virus capsids [3]. This chapter will give a brief history of mass spectrometry, focusing on developments in instrumentation relevant to this thesis.

The inventor of the mass spectrometer was the English physicist Joseph John Thomson, usually referred to as J.J. Thomson. Following his early career as a theoretical physicist, he was awarded a professorship at the University of Cambridge as an experimental physicist. He investigated the movement of electricity through gas, in particular cathode rays. The research determined that a cathode ray could be deflected using magnets, and this information was used to posit that the cathode rays were made up of particles more than 1,000 times lighter than a hydrogen atom by calculating its z/m ratio. This revelation in 1897 was the discovery of the first subatomic particle, the electron and for which he was awarded the Nobel Prize in Physics in 1906.

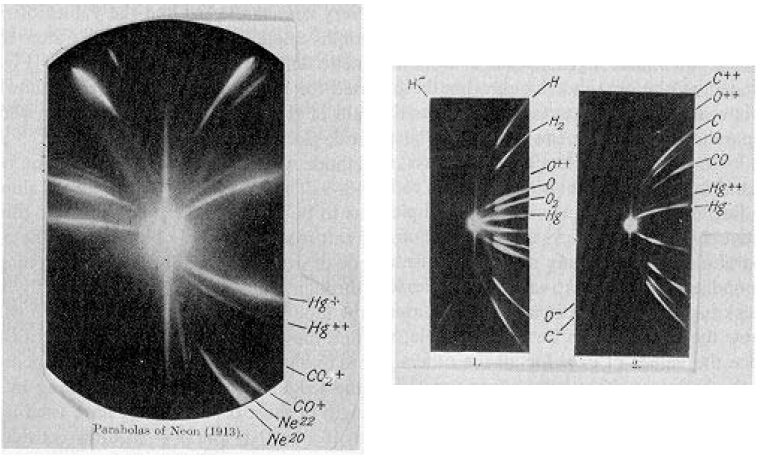

Figure 1.1: Parabola spectrographs. Figure adapted from [4].

The work was extended to look at positive ions in an instrument called the parabola spectrograph. By deflecting ions using magnetic and electric fields onto photographic plates, spectra like those shown in Figure 1.1 were acquired. Thomson postulated that any ion would have a different parabolic trace provided that they had different m/z values. The discovery of an additional trace at 22 Da appeared to be neon, as had been suggested by Thomson. By building the first mass spectrometer his protege Francis William Aston, in 1919 [5],was able to label these as isotopes and classified several other isotopes [5]. He also measured the mass of as many elements as he could, which lead to him defining the whole number rule that states that every elements mass is an integer multiple of the mass of a hydrogen atom [6]. These two discoveries led to him being awarded the Nobel Prize for Chemistry in 1922.

In the early days mass spectrometers were mainly used by physicists on subjects such as determining features of atoms. Alfred Nier was a pioneer who popularised the instrumentation outside the field of physics. He was an electrical engineer and physicist by training and used these attributes to create many mass spectrometers including the Nier-Johnson mass spectrometer. His promotion of mass spectrometry included using his instruments to filter carbon allowing for the collection of enriched ^{13}C for use in biological applications. His most famous contribution however was to the Manhattan Project. It was known that one of the isotopes of uranium was fissile, but it was not yet known which. In 1939 he was recruited by Enrico Fermi to

separate the two isotopes; work which led to the discovery that ^{235}U was capable of nuclear fission thereby bringing in the nuclear age [7]. Throughout the Manhattan Project mass spectrometers he designed were used to assess the purity of enriched uranium.

Important developments continued to be made in mass spectrometry instrumentation in the late 1940s and 1950s. The time-of-flight (ToF) mass analyser was devised in 1946 by Stephens [8], it was then built by Cameron and Eggers with the results published in 1948 [9]. The ToF mass analyser was based on simple physics, accelerating ions by their charge and determining the mass by the velocity achieved over the distance of a vacuum tube. Modern day implementations of the instrumentation are now capable of analysing very high mass ions. The linear quadrupole mass analyser was introduced in 1953 by Paul and Steinwedel [10], and can be used as an ion guide or a mass filter [11]. These two components can be combined to create a Q-ToF. This instrumentation has become very popular in biological and analytical chemistry applications, including being linked to liquid chromatography for proteomics studies and for studying intact proteins.

The analysis of small molecules had become routine before the 1980s but it was still not possible to analyse large biological molecules like proteins. This was due to ionisation techniques of the time which consisted of bombarding the analyte in the gas-phase with charged particles. This technique deposited too much energy onto proteins causing them to fragment. Later techniques such as fast atom bombardment were able to ionise the analyte when not in the gas-phase which was an improvement, but was very inefficient and still would not work for larger biomolecules [12].

In 1985 Koichi Tanaka filed a patent for a new method of ionisation [13]. His method involved mixing the analyte with glycerol and metal powder, followed by ionisation with a laser, and was called soft laser desorption (SLD). He found that the inclusion of the metal reduced the transfer of energy to the analyte and allowed for larger molecules to be analysed. This work was presented at the Annual Conference of the Mass Spectrometry Society of Japan and led to him being awarded the Nobel Prize for Chemistry in 2002 with John Fenn [14]. This method, however, is presently rarely used for biomolecule analysis.

A variant, invented alongside SLD, was MALDI (matrix assisted laser desorption ionisation) which was developed by Karas and Hillenkamp [15]. By 1988 Karas and Hillenkamp were able to analyse molecules over 10,000 Da [16] and MALDI has widespread usage for analysing biomolecules to this day.

In 1968 Dole *et al.* introduced the concept of electrospray ionisation (ESI) by ionising large polystyrene molecules up to 400 kDa, a feat which was previously not possible due to the alternative method at the time, evaporation through heating, degraded large molecules [17]. Several years later John Fenn introduced the use of electrospray ionisation mass spectrometry for use with large biological molecules, at first with small organic molecules in 1984 [18], this was then extended to proteins up to 40 kDa in 1988 [19] and by 1989 he had successfully analysed the dimeric species of bovine serum albumin (130 kDa) [20, 21]. This technique ionised the analyte while in solution phase, which allowed for very gentle ionisation and was ideally suited for analysis of intact proteins. Additionally ESI created multiply charged ions and this allowed high mass analytes to be analysed using mass analysers with limited m/z ranges. The technique is now very common, and all spectra in this thesis were obtained using nano-ESI, a variation of the ESI method which will be described in Section 2.1.

Chapter 2

Mass spectrometry

This chapter introduces the components of the Waters Synapt G1 mass spectrometer that was used during the Ph.D. The instrument consists of a nanoelectrospray ionisation source, a quadrupole mass analyser, a travelling wave ion mobility separation device (Covered in Chapter 3), a time-of-flight mass analyser and an ion detection device. An example mass spectrum acquired on the instrument is shown in Figure 2.1. This is followed by examples of mass spectrometry experiments that are relevant to this thesis and the challenges raised by each.

2.1 Electrospray ionisation

Samples for ESI are loaded into a metal needle or a gold coated glass capillary and there is a potential applied between the needle and the entrance to the mass spectrometer (Figure 2.2A). In positive ion mode, which is normally used for protein analysis and is used exclusively in this thesis, the positive voltage is applied to the needle, and the negative voltage is applied to the counter-electrode in the instrument. This leads to a build up of charge in the form of protons at the tip of the needle. At first this leads to a spherical body of analyte solution forming on the tip of the needle, and as the voltage

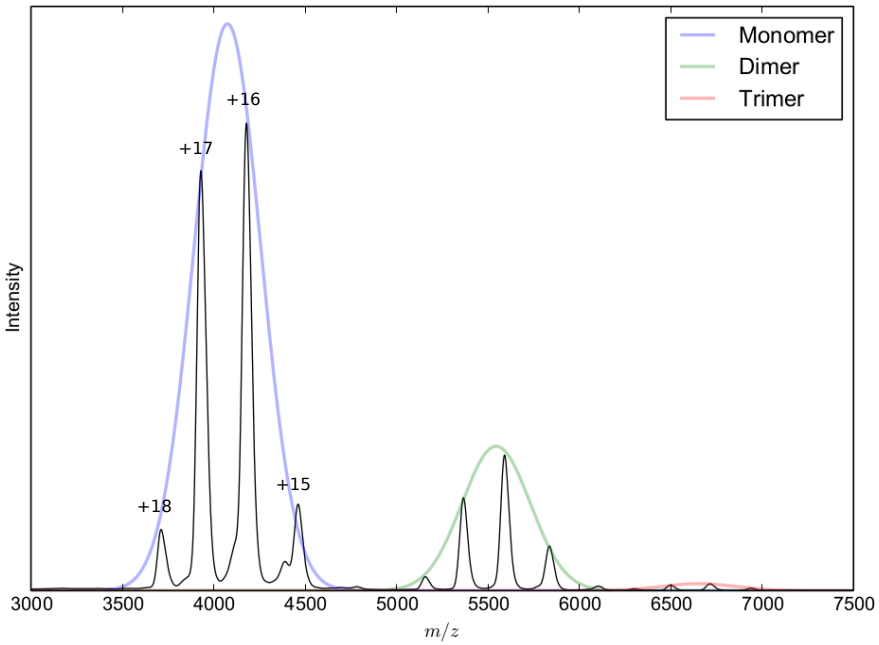

Figure 2.1: Example nESI mass spectrum, containing the monomer, dimer and trimer of human serum albumin. Gaussian distributions are fit to the peak tops of each molecular species and are displayed as faded lines. Additionally the charge states of the monomer peaks have been labelled.

increases the shape distorts. Finally when the surface tension limit is reached the solution forms a Taylor cone [1], and emanates a spray of fine droplets which are drawn towards the counter-electrode (Figure 2.2B). The flow of protons out of the capillary is offset by the transport of electrons across the voltage source to the counter-electrode. This maintains the charge difference and allows electrospray to continue.

The droplets leaving the capillary are large and as the molecule travels, the buffer evaporates reducing the size of the droplet. The concentration of positive charges inside the droplet increases as the volume decreases and the coulombic repulsion makes the droplet unstable and eventually causes droplet fission. The Rayleigh limit, shown in Equation (2.1) [3], defines the point at which the coulombic repulsion overcomes the surface tension. This is reached when q (charge of the droplet) becomes greater than the Rayleigh limit, q_r, other terms in the equation are: ϵ_0 the electrical permittivity of a vacuum, γ and R are the droplet surface tension and radius respectively. After droplet

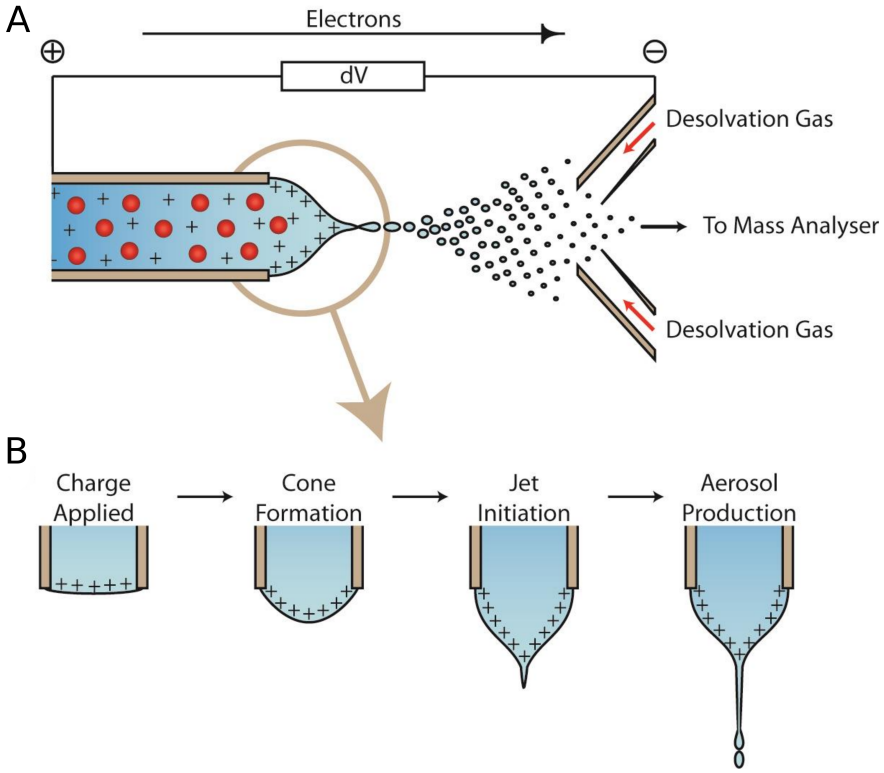

Figure 2.2: Schematic diagrams of the electrospray ionisation process. The formation of droplets at the tip of an electrospray needle (A) and the formation of a Taylor cone (B). Figure adapted from [2].

fission, the daughter droplets undergo further solvent evaporation, leading to further fission events, this process continues iteratively until very small droplets remain which contain a single analyte molecule [4].

$$q_r = 8\pi\sqrt{\epsilon_0 \gamma R^3} \tag{2.1}$$

The final stage of ESI is still not fully understood. There are three main theories about what happens to produce the final analyte ion which is analysed by the mass spectrometer. The first model proposed was in the original electrospray ionisation publication by Dole *et al.* [5], the theory is known as the charged residue model (CRM) and it states that the fission-evaporation cycle continues until the analyte, adducted with the remaining charges, is left in a dehydrated state. The second theory, stated by Iribarne and Thomson in 1976 [6], is known as the ion evaporation model (IEM). In an electrospray

droplet the charges are at the highest concentration at the surface, due to coulombic repulsion. The IEM suggests that the analyte would be forced through the surface of the droplet, during this process the analyte becomes the least electrostatically repulsive portion of the droplet, causing it to pick up several charges, and be ejected in a dehydrated state.

The desolvation of small molecular weight ions has been attributed to the IEM [6], with larger compact ions such as globular proteins following the CRM [7, 8]. A recent molecular dynamics study has revealed a third method of ion desolvation called the chain ejection model (CEM) [9]. The CEM is thought to explain the desolvation of proteins that are unfolded in solution, and is a similar process to the IEM. The hydrophobic core of unfolded proteins is exposed to solvent and so the protein quickly migrates to the surface of a droplet. One end of the protein chain will start to exit the droplet, due to charge repulsion. The chain is then steadily drawn out of the droplet, whilst picking up charges, until completely free of the original droplet [9].

2.1.1 Interpreting ESI mass spectra

Electrospray ionisation usually results in multiply charged ions [10], and an example mass spectrum is shown in Figure 2.1. The charge of protein ions is due to protonation of basic amino acid side chains [11]. The protonation process results in a mass spectrum with multiple peaks with intensities similar to that of a Gaussian distribution [12]. The explanation for this is the central limit theorem, a mass spectrum is the summation of a large number of ion detections, and the number of protons adducted to an analyte are consecutive, have an average value, and a certain variance which leads to the Gaussian-like distribution.

As analytes in an ESI spectrum have multiple m/z values, it can make mass analysis more complicated than a technique like MALDI, which typically only produces singly charged ions. The mass (M) can be determined by the number of charges (z) and m/z value (m_x), and a straight forward calculation is described here. Two adjacent peaks originating from the same analyte will have a charge difference of 1. This allows Equations 2.2 & 2.3 to be true (H^+

is the mass of a proton).

$$m_1 = \frac{M + z \cdot \mathrm{H}^+}{z} \tag{2.2}$$

$$m_2 = \frac{M + (z+1) \cdot \mathrm{H}^+}{(z+1)} \tag{2.3}$$

$$n = \frac{m_2 - \mathrm{H}^+}{m_1 - m_2} \tag{2.4}$$

Solving the simultaneous equations for charge gives Equation 2.4, which determines the charge for Equation 2.2. Equation 2.2 can be rearranged to give Equation 2.5 and from this, the mass of the ion can be calculated.

$$M = z(m_1 - \mathrm{H}^+) \tag{2.5}$$

2.1.2 Nanoelectrospray ionisation

Electrospray ionisation was improved with the developments of microelectrospray ionisation (μESI) [13, 14] and then nanoelectrospray ionisation (nESI) [15]. The prefixes of these new ionisation techniques were based on the sample flow rate during experiments. ESI experiments typically have a flow rate of 1-20 μl/min whereas the rate for nESI is only $\sim$20 nl/min, this difference is due to a difference in the orifice size of the electrospray needle, with inner diameters of 100 μm for ES [16] and 1-2 μm for nESI.

The reduced flow rate of nESI meant that the droplets leaving the needle were considerably smaller, with less solvent, the result of which was that the droplet had to undergo less desolvation in order to accurately assess the mass of proteins in the gas-phase. The reduction in solvent volume in a droplet results in a lower total number of salt ions, leading to a lower concentration of salts as the droplet undergoes desolvation. This results in less cation adduction and thinner peaks in a mass spectrum [17]. This also facilitated a

shorter necessary distance between the needle tip and the entrance of the mass spectrometer, and concomitantly reduced the necessary voltage difference for ionisation resulting in less energy being transferred to the analyte molecule, and achieving 'softer ionisation'.

The large droplet sizes of electrospray ionisation preclude the use of aqueous solvents, and for this reason organic solvents with their low boiling temperatures were used instead. An example solvent mixture for the ESI analysis of proteins (Cytochrome *c*) would be 45 % methanol, 45 % acetonitrile, 10 % water and 167 ppm trifluoroacetic acid [18].

The reduction in the required energy for ionisation afforded by nESI allowed the use of the volatile salts, ammonium acetate or ammonium bicarbonate dissolved in water [19]. Volatile salts evaporate readily during the desolvation process, this leaves the analyte with considerably less adducts than when using non-volatile buffer salts (e.g. NaCl). Ammonium acetate is primarily used [19], and the salt aids in the ionisation process as it competes for adduct formation with basic side chain residues with cations such as Na^+, this is due to the process whereby NH_4^+ ions protonate the sidechains leaving volatile ammonia that subsequently evaporates [20]. Additionally ammonium acetate has a similar ionic strength to sodium chloride, meaning that they can be used in place of NaCl in order to retain physiological salt levels.

The organic solvents and energy conditions of electrospray ionisation make it difficult to analyse proteins in their native conformation as organic solvents disrupt hydrophobic interactions causing the protein to denature [21]. The use of aqueous buffers at physiological ionic strengths, meant that nESI experiments were able to analyse the native-like structure of larger proteins [22] as well as analysing small proteins more easily. This has led to nESI mass spectrometry becoming an essential tool in the analysis of protein-ligand [23], protein-protein [24] and protein structural information [25].

2.1.3 Mass spectrometry peak shapes

An unadducted ion mass spectrum peak is Gaussian when analysed using a Waters Synapt G1 [26]. Proteins, however, have additional peaks corresponding to adduction as well. This can be seen clearly in a mass spectrum of the small protein β-lactoglobulin in Figure 2.3A. The first peak, indicated by the arrow, is the protein in its unadducted state, and the theoretical and experimental masses as calculated using this peak are very accurate. The additional peaks at higher m/z values indicate cation adduction, normally a mixture of sodium and potassium ions. These peaks are not baseline separated, and so the intensity of several combinations of adduct peaks combine to form the reducing intensity tail of the charge state peak.

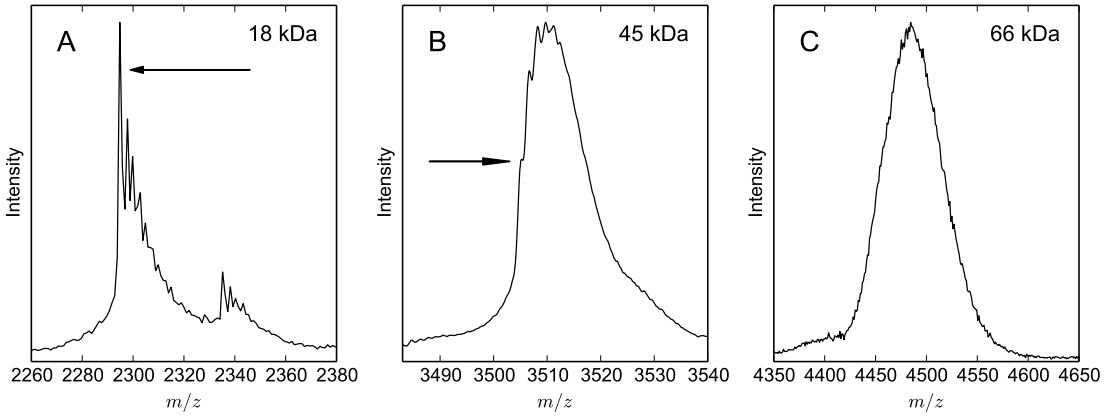

Figure 2.3: Comparison of different m/z value peak shapes. Arrows point to unadducted protein peak for β-lactoglobulin monomer (A), α_1-antitrypsin (B) and bovine serum albumin (C).

Figure 2.3B shows a charge state peak of the larger protein, α_1-antitrypsin, once again with the arrow indicating the unadducted ion. With this protein, however, the unadducted ion is only detected as a peak and is difficult to distinguish from the rest of the charge state. The highest point of the peak has now moved to a higher m/z than the theoretically calculated value. The summing of many adduct Gaussian shaped peaks causes baseline lift and forms the new peak shape typical of proteins, with a sharp increase in intensity at the lower m/z values, followed by a rounding off and slower decline in intensity towards the upper m/z range of the peak.

With larger proteins it is often difficult to distinguish any of the individual ion or ion-adduct peaks from the overall charge state peak. An example of this is shown in Figure 2.3. Experiments can be designed to try to minimise this effect, such as denaturing mass spectrometry which, will be discussed further in Section 2.1.5.

There are several reasons for this deterioration in data quality. One such reason is that larger proteins are able to be adducted more, resulting in the larger number of adduct peaks. The mass resolution of the Waters Synapt G1 (and most mass spectrometers), reduces the higher the m/z value, meaning that it is more difficult to separate peaks. This feature will be discussed in more detail in Section 2.2.

The spacing of adduct peaks also changes for larger proteins. Typically larger proteins have a higher number of charges, in the example shown the most abundant charge states were used and they were +8 for β-lactoglobulin, +13 for α_1-antitrypsin and +15 for bovine serum albumin. A sodium adduct peak adds a mass of 23 Da to the mass of the protein. For a +7 protein, the m/z difference between the unadducted ion and an ion with a single sodium adduct is 3.3 ($m/z = 23/7$), whereas for a +15 protein, the m/z difference would only be 1.5. This increases the extent of the overlap between adducted and unadducted peaks, making analysis more difficult.

In order to analyse higher mass protein ions, the ions need to be cooled after exiting the spray tip. This is achieved by applying a counter current drying gas that helps with desolvation [27]. Additionally the buffer gases within the mass spectrometer, in regions such as the collision cell, should also have higher gas pressures than when analysing smaller proteins [28]. Higher mass ions require higher collision energy voltages in order to achieve the same acceleration, and so this also needs to be adjusted for when analysing larger proteins.

2.1.4 Non-covalent mass spectrometry

Non-covalent mass spectrometry aims to analyse proteins or other analytes without disrupting non-covalent interactions, it is also sometimes referred to as native mass spectrometry. Some examples of uses for these experiments are to analyse protein complexes without dissociating subunits and analysing protein ligand interactions.

In order to analyse proteins without disrupting the protein structure, careful consideration must be taken to the conditions of the sample prior to the introduction into the mass spectrometer as well as to the settings of mass spectrometer itself. Proteins and protein complexes dissolved in organic solvents can have structural elements broken down due to the reduced entropy maintaining the hydrophobic core of the protein. This can lead to structural re-arrangements which cause the dissociation of ligands or subunits. Acidic buffers can also negatively affect structure, as native protein structure is dependent on physiological pH levels. In order to maintain the non-covalent interactions, it is therefore beneficial to use aqueous buffers with physiological pH and ionic strengths to best preserve protein structure. For these experiments, an appropriate example solution for a protein analyte would be 150 mM ammonium acetate (which has similar ionic strength to sodium chloride) at pH 7. An example spectrum can be seen in Figure 2.1.

The protonation of globular and disordered proteins varies due to the spatial arrangement of basic residues (the site of protonation). In globular proteins these side chains are close to each other and charge repulsion between protonated side chains will reduce the level of protonation. This is in comparison to unfolded proteins where the side chains will be more spread out, lessening the coulombic repulsion. A measure for this is the solvent accessible surface area (SASA), which has been shown to be directly related to the level of protonation and ultimately, charge state distributions [29].

Due to the changes in protonation, the extent of protein folding can be seen, to some extent, in a mass spectrum. More globular proteins have a narrower charge state distribution (number of charge state peaks), in comparison to unfolded proteins with disrupted non-covalent interactions, as can be seen in

Figure 2.4. Myoglobin has been analysed using ammonium acetate at ph 7.5 and acidic organic buffers. The protein analysed in aqueous buffer has a far tighter charge state distribution, and the haeme group is still attached. This is in comparison to methanol and acetic acid, which shows many charge states as well as having the haeme group, which is natively non-covalently bound, dissociated (the pronounced low m/z peak).

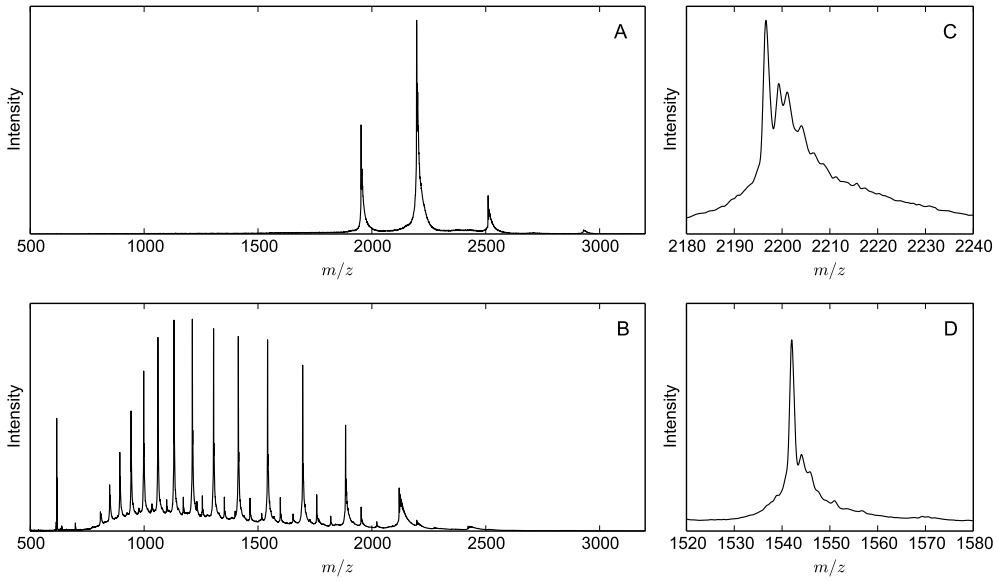

Figure 2.4: Nanoelectrospray ionisation mass spectra of myoglobin in different sample buffers. (A) Non-denaturing conditions (ammonium acetate at pH 7.5), and denaturing buffer (methanol and acetic acid) (B). An enlarged view is also shown for a charge state peak under non-denaturing conditions (C) and denaturing conditions (D).

After it has been ensured that non-covalent interactions have been maintained and proteins have retained their structure, the mass spectrometer settings should be assessed. The Waters Synapt G1 mass spectrometer has several user defined settings, and some of the most important ones to non-covalent interactions will be discussed. The capillary voltage and cone voltages determine the voltage difference which creates the Taylor cone and begins the electrospray process. These should be kept to a minimum so as not to transfer too much energy into the analyte, which could potentially break hydrogen bonds. The added acceleration which high capillary and cone voltages would create would also affect the next important region, the T-wave device (discussed in more depth in Chapter 3). These are gas filled chambers and

the user can set the voltage difference between the entrance and exit of each chamber. These voltage differences should also be kept to a minimum. This is because increasing the voltage, causes increased acceleration of the analyte ions, which results in more violent collisions with the buffer gas and due to friction this causes the ions to heat and can potentially break non-covalent interactions. If these voltages have been kept low enough, non-covalent interactions should be maintained throughout the mass spectrometry experiment. Ion mobility mass spectrometry has been used to show that this is true and the subject is discussed in more depth in Section 3.4.

Non-covalent mass spectrometry is necessary for determining the mass of protein complexes together because if a subunit dissociates the recorded mass will be incorrect.

Determining binding stability

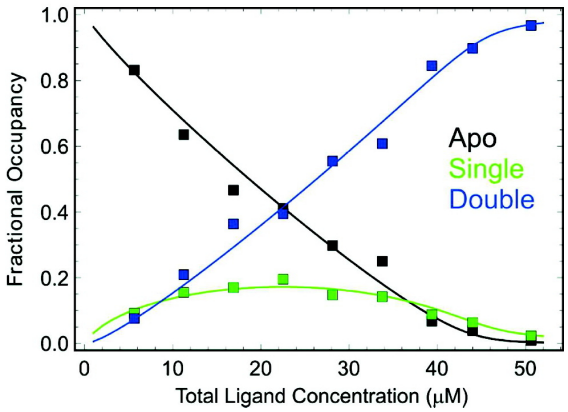

Figure 2.5: Analysis of the interaction between GyrA59 and SD8. The fractional occupancy is calculated from the percentage of apo, single and double bound peak intensities in mass spectra. The abundance of each species is recorded for different concentrations of ligands (square markers) and the models for calculating K_d values are shown as the fitted curves. Figure reproduced from [30].

Protein-ligand experiments that determine binding stability in terms of calculating the dissociation constant $(K_d)$of an interaction are well suited to non-covalent mass spectrometry analysis. As previously mentioned it is vital that the solution conditions and mass spectrometer parameters do not disrupt hydrogen bonding, which could lead to the dissociation of the ligand. A good example of the benefits of mass spectrometry for these types of experiments

is a 2011 study by Edwards *et al.* [30]. The study analysed the interaction between GyrA59 and SD8. It was found that the GyrA59 protein associates as a dimer, with one ligand bound to each subunit. It was then shown that the binding was cooperative by altering the concentration of the ligand and measuring the mass spectrometry intensities of each bound state as seen in Figure 2.5.

A mathematical model was developed to describe the binding interaction and the results are shown as the fitted curves in Figure 2.5 using the abundance of each bound state. Non-covalent mass spectrometry was ideal for analysing this interaction, due to the K_d calculations being complicated by stoichiometries, which can cause issues for other analytical tools [30].

Subunit exchange

As well as studying equilibrium state interactions such as determining K_d values, non-covalent mass spectrometry can be used to analyse protein dynamics, in particular, subunit exchange. These are time-resolved experiments that monitor the rate at which protein subunits spontaneously exchange in solution. For these types of experiments to be analysed by mass spectrometry, there needs to be a mass difference between the subunits. The experiment is carried out by mixing the two subunits in solution, then repeatedly acquiring mass spectra of the sample at a given time interval. Each combination of subunits will have a different mass, thereby resulting in a separate peak in the mass spectrum. An example would be a heterodimer with subunits A and B, this will result in mass spectrum peaks associated with 2A, 2B and AB. The abundance of these subunits can then be analysed using the heights of these peaks and can be plotted (as in 2.6).

A particularly interesting example was published in 2011, which involved monitoring a homodimer [31]. This was achieved using nitrogen labelling, with ^{14}N and ^{15}N protein being used.

The protein, glucosamine-6-phosphate synthase (GlmS) had been well studied by X-ray crystallography, but had not been investigated for subunit exchange. GlmS converts D-fructose-6-phosphate (Fru6P) into D-glucosamine-6-phosphate (GlcN6P) and the interaction uses L-glutamine as a nitrogen

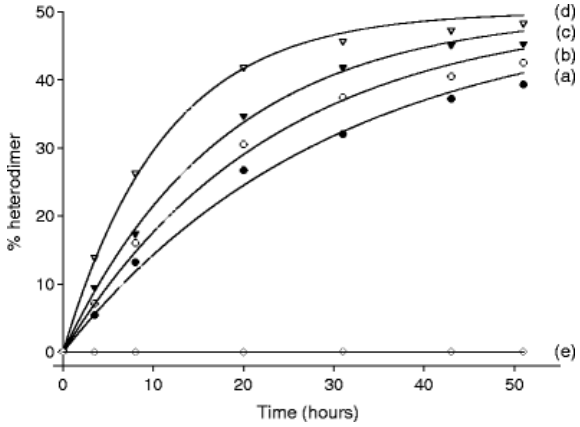

Figure 2.6: Monitoring the rate of subunit exhange of GlmS, with extent of exchange reported as the percentage of $^{14}N/^{15}N$. The results are shown with 4 μ M GlmS alone (a), as well as in the presence of glutamine or glutamate at 5 μ M (b), 10 μ M (c) and 100 μ M (d). The rate of subunit exchange is also reported for 20 μ M GlmS with 100 μ M of Fru6P or GlcN6P (e). Figure reproduced from [31].

donor, which is then converted to glutamate.

The researchers monitored the rate of subunit exchange in the presence of both L-glutamine and Fru6P. The sample monitored would contain both the initially added molecules and the reaction products, and the results are shown in Figure2.6. It was shown that the complex naturally undergoes subunit exchange, and that this rate is increased as the concentration of the nitrogen donor is increased. The presence of the substrate, however, stopped subunit exchange all together [31].

2.1.5 Denaturing mass spectrometry

It is sometimes advantageous to intentionally unfold a protein in solution. This can be achieved by varying buffer conditions and some examples will be given here.

When analysing an unknown protein complex it is beneficial to break the complex into its component parts to determine the subunits. Unfolding the protein causes disruption of the interface between protein subunits, causing them to separate. This can be achieved by using acidic organic solvents as previously described. Subunit dissociation can also be achieved using collision

induced dissociation (CID) which is discussed in Section 2.2.3.

The assembly pathway of protein complexes can also be analysed using denaturing conditions. The principle of this experiment is to increase the harshness of the solution by increasing either organic solvent or ammonium acetate concentration, and to observe which subunit dissociates first. An example of this type of experiment is shown in Figure 2.7. As the ammonium acetate concentration is increased the abundance of dimers and momomers also increases with no trimeric species observed. This shows that the assembly pathway of the hexamer is through the association of three dimers, and not consecutive adding of monomers or the association of pairs of timers [32].

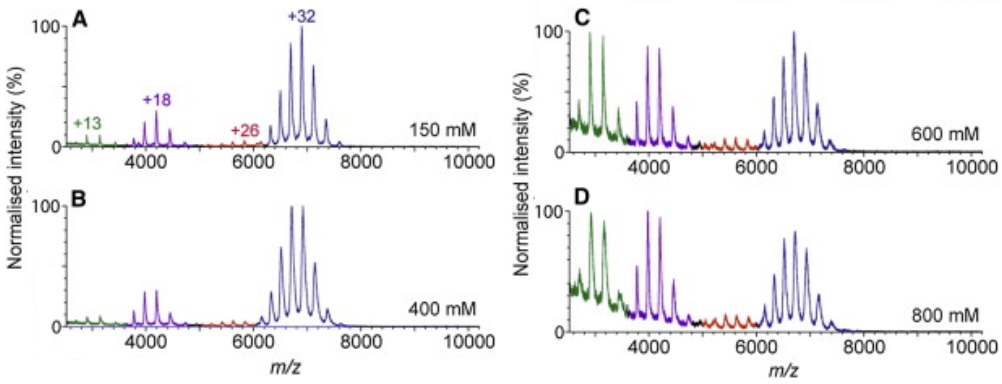

Figure 2.7: Mass spectra of ExsG at varying ammonium acetate concentrations. The hexameric (blue), tetrameric (red), dimeric (purple) and monomeric (green) protein has been coloured. Spectra are shown for 150 mM (A), 400 mM (B), 600 mM (C) and 800 mM (D) ammonium acetate. Figure adapted from [32].

Another reason for carrying out solution denaturing experiments is to determine highly accurate protein masses. As previously mentioned, proteins are adducted by salts during the desolvation process. This results in a protein (M) charge state peak being composed of several ion mass peaks (e.g. $M + Na$ and $M + 2Na$). This makes it difficult to determine the mass of a protein as the peaks are an aggregate of several different peaks of varying masses.

By unfolding the protein in solution the level of adduction can be greatly reduced, due to the final stage of desolvation. Globular proteins follow the charged residue model (CRM), where the final droplet evaporates with the protein still within it, which causes the remaining non-volatile salts from the

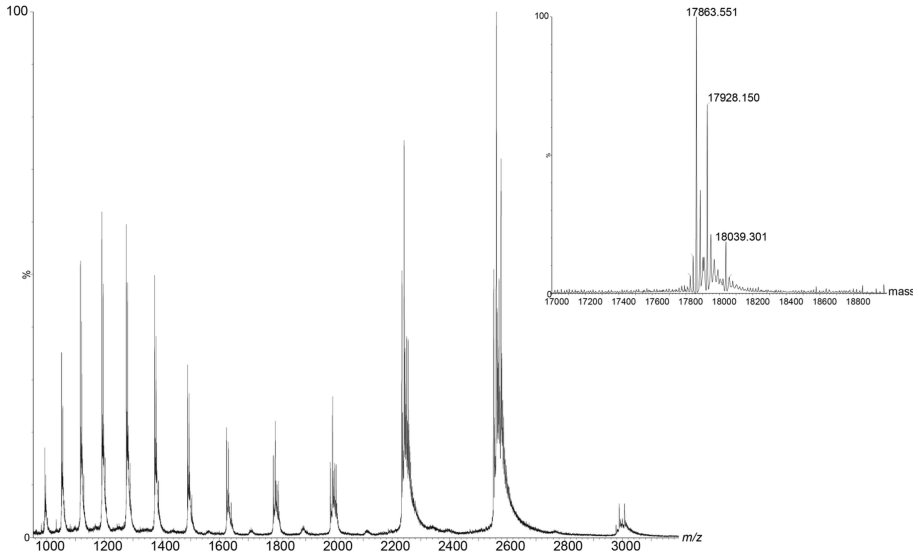

Figure 2.8: Mass spectrum (main) and MaxEnt spectrum (inset) of denatured AioB. The MaxEnt spectrum shows the distribution of masses in the unfolded region of the mass spectrum (<2,200 m/z). The theoretical mass of AioB without the disulphide bond is 17863.76 Da and with the bond is 17861.75 Da. The mass determined by denaturing MS was 17863.551, showing that the bond had not formed. Figure reproduced from [33].

buffer to be adducted onto the protein. Unfolded proteins follow the chain ejection model (CEM), whereby they migrate to the surface of the droplet and are ejected out, which results in few salt adduction events.

An example of this type of experiment is shown in Figure 2.8. In this study it was necessary to determine whether an 18 kDa protein (AioB) had formed a disulphide bond using mass spectrometry. The reaction is shown in Equation 2.6 and shows that the mass difference is 2 Da (H_2).

$$2\text{Cys} - \text{SH} \rightarrow \text{Cys} - \text{S} - \text{S} - \text{Cys} + 2\text{H}^+ + 2e^- \tag{2.6}$$

In order to achieve mass accuracy to this level, the protein must be analysed without any adducts and so a solution denaturing experiment was conducted. The result was that the mass was calculated to within 0.25 Da of the calculated value [33].

2.2 Mass analysis and ion detection

Different types of mass analysers have different characteristics and so are useful for different types of analysis. The way in which they determine m/z can be different, scanning analysers such as the quadrupole scan regions of the mass spectrum in separate acquisitions, whereas a continuous m/z analyser such as time-of-flight (ToF) mass analysers acquire a full m/z range simultaneously.

There are three important characteristics about the data acquired by each instrumentation, mass range limit, mass accuracy and mass resolution (and resolving power). The mass range limit refers to the range of possible m/z values that a mass analyser can detect. Quadrupoles have limited m/z ranges, with the standard Waters Synapt G1 quadrupole only being able to acquire up to 8,000 m/z, whereas ToFs have theoretically unlimited m/z ranges [34].

The mass accuracy of a mass analyser is determined by calculating the difference in the theoretical and experimental mass and is measured in parts per million [35]. The ToF instrument used in this thesis is calibrated using a solution of caesium iodide, which forms many different sized crystals, and the target accuracy is better than 10 pm.

The resolving power (R) is also an important feature of mass analysers. It shows how close in m/z two ions can be and still be separated by the instrument. Through the years, the resolving power of mass analysers have improved greatly with Aston's parabola spectrograph achieving $R = 13$, whereas the time-of-flight mass analyser in the Waters Synapt G1 has a resolving power of 10,000.

Mass resolution (Δm)is typically measured as the full width half maximum (FWHM) of a particular ion, as seen in Figure 2.9A. The relationship is described in Equation 2.7, where m_1 is the lower m/z boundary of the peak at 50 % of the peak height and m_2 is the upper boundary.

$$\Delta m = m_2 - m_1 \tag{2.7}$$

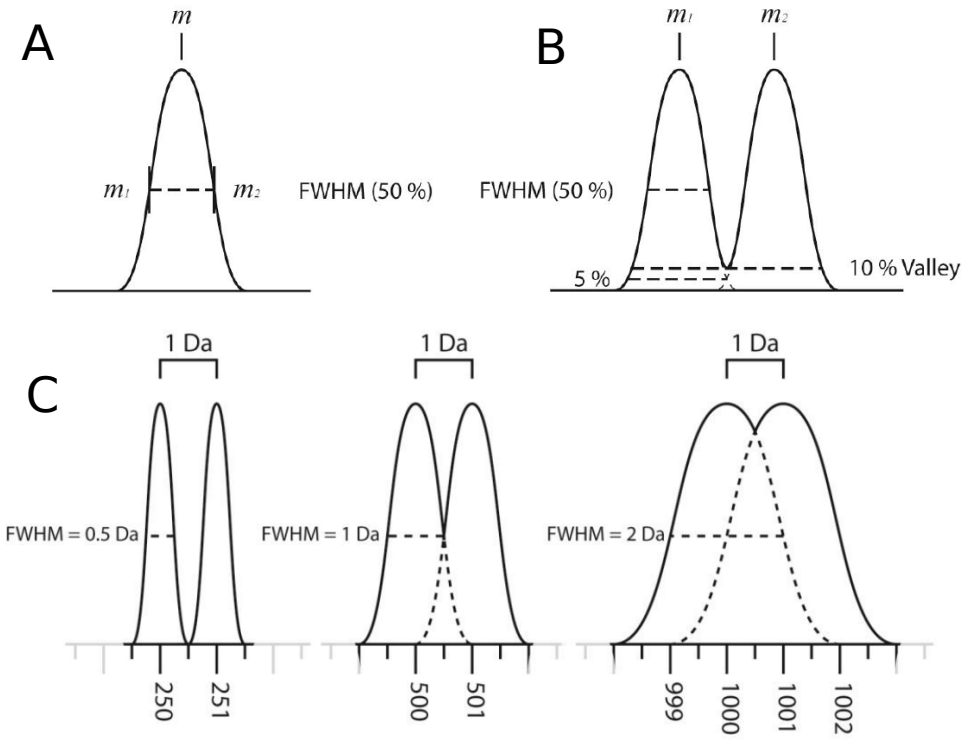

Figure 2.9: Figure showing the variables involved in the calculation of mass resolution (Δm) and resolving power (R). (A) the determination of resolution is calculated as per Equation 2.7. (B) representation of alternate valley heights in the calculation of resolution, the example shows the separation required for a 10 % valley. (C) demonstrates how the m/z value affects resolving power. At a given resolving power peaks at a lower m/z range are more separated than peaks at a higher mass range, which shows that, in order to analyse higher m/z ions, the mass analyser requires a higher resolving power to separate peaks. Figure adapted from [2].

This means that if two ion peaks at the same intensity were separated by Δm m/z, the trough between the peaks would be at 50 % of the peak height, as seen in the centre panel of Figure 2.9C.

The resolving power of a mass analyser additionally includes the m/z region (m) at which a particular resolution is achieved, and the relationship is shown in Equation 2.8.

$$R = \frac{m}{\Delta m} \tag{2.8}$$

Sometimes rather than analysing resolution and resolving power as the

ability to separate peaks to half height, the value can be reported in terms of other heights such as the 10 % valley. With this definition it is necessary for peaks to be separated to an extent that the trough between them is at 10 % of the total peak height as shown in Figure 2.9B [36].

2.2.1 Linear quadrupole mass analyser

Linear quadrupoles consist of 4 parallel metal rods each with a circular or hyperbolic cross section (Figure 2.10).

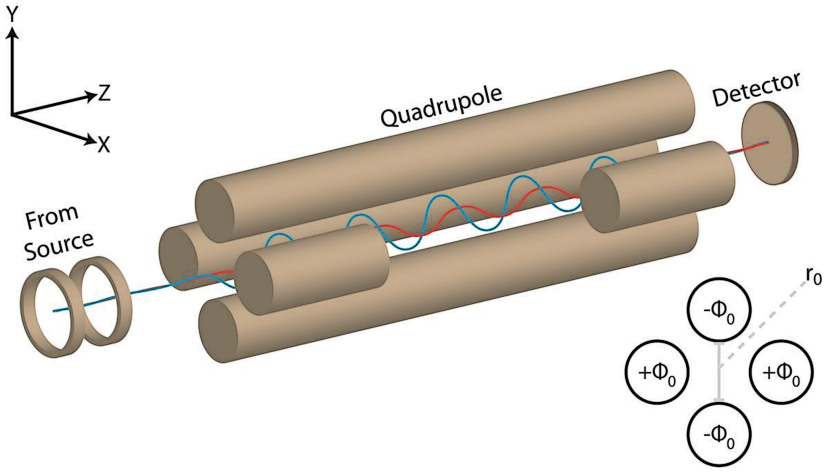

Figure 2.10: Illustration of the workings of a quadrupole mass analyser. Blue and red lines indicate the path of ions, caused by the alternating electric field applied by the quadrupole. Figure adapted from [2].

The rods have a direct potential (U) and a radio frequency (RF) voltage (amplitude is V) applied to them. The alternating pairs of rods are oppositely charged as described in Equations 2.9,2.10 where the angular frequency (rotation rate) is shown as ω. The RF potential frequency (v) can then be calculated from the angular frequency using Equation 2.11.

$$\Phi_0 = +(U - V\cos\omega t) \tag{2.9}$$

$$-\Phi_0 = -(U - V\cos\omega t) \tag{2.10}$$

$$\omega = 2\pi v \tag{2.11}$$

Ions travel along the z axis at a constant velocity, but the combination of RF and direct current voltages cause the ions to undergo complex oscillations. Particular combinations of RF and direct current voltages are required for the ion to maintain a stable trajectory through the quadrupole. By altering these voltages particular m/z ions can be selected, whereas other ions will collide with the rods and leave the quadrupole.

Quadrupoles can also be used as ion guides, which has the effect of high mass filtering. This is achieved by running the quadrupole in RF only mode. The maximum m/z (M_{max}) for a given quadrupole is given by Equation 2.12 where r is the radius of the quadrupole.

$$M_{max} = \frac{7 \cdot 10^6 V}{v^2 r^2} \tag{2.12}$$

In the instrument used for the experiments described in this thesis, both 8,000 and 32,000 m/z upper mass limit quadrupoles were used. The 32,000 quadrupole has a modified RF voltage generator allowing for lower frequency radio wave generation allowing the analysis of higher masses.

2.2.2 Time-of-flight mass analyser

In order to determine the m/z ratios of ions, packets of ions are accelerated, at a common energy per unit charge, into a time-of-flight (ToF) mass analyser. The ions then traverse the field-free region, where they undergo no further acceleration. Given two ions with different mass but the same charge, the velocity through the field-free region is higher for the lower mass ion. The m/z ratio is then calculated from the time taken for ions to travel from the start of the field-free region until reaching an ion detector at the end of the ToF.

The electrical potential energy (E) gained from the acceleration stage (prior

to entering the field-free region) can be calculated from the potential (V) and the charge q. In mass spectrometry charge is referred to in terms of the charge of an electron (e) multiplied by charge state (z) of the ion, and so the potential energy of an ion can be calculated using Equation 2.13. The potential energy is converted to kinetic energy, which relates to the mass (m) and velocity in the field-free region (v) as shown in Equation 2.14.

$$E = zeV \tag{2.13}$$

$$E = \frac{mv^2}{2} \tag{2.14}$$

Combining Equations 2.14 and 2.13 produces Equation 2.15 which allows the calculation of velocity directly. As the ion moves through the flight tube without obstruction, the velocity is constant following the acceleration step. This means the time taken (t) for an ion to traverse the length of the flight tube (d) can be calculated with Equation 2.16.

$$v = \sqrt{\frac{2zeV}{m}} \tag{2.15}$$

$$t = \frac{d}{v} \tag{2.16}$$

Substituting v into the equation for time, results in Equation 2.17. The charge of an electron, length of the flight tube and electric potential across the tube remain constant. This means that changes in the m/z value of an ion will directly affect its time to traverse the flight tube; increasing z will reduce the time and increasing m increases the time.

$$t^2 = \frac{m}{z} \cdot \left(\frac{d^2}{2eV}\right) \tag{2.17}$$

Limiting factors in mass resolution

The mass resolution of ToF mass spectrometers when analysing large analytes such as proteins is now described. The measurement of the time taken to traverse the field-free region and reach the detector has a certain precision that depends on the instrument and which cannot be improved by the user. nESI analysis of proteins produce peaks which typically have several adducted cations (in positive ion mode). The broadening of peaks makes it difficult to separate peaks, and causes overlapping and masked peaks in mass spectra. In positive ion mode the adduction is typically from sodium or potassium ions. It is therefore essential to minimise the concentration of these ions and their salts in the sample solution in order to produce well defined peaks. This can be achieved by using buffer exchange before mass spectrometry analysis and replacing the initial buffer with an appropriate mass spectrometry buffer, such as 150 mM ammonium acetate. This is demonstrated in Figure 2.11, when in sodium phosphate buffer the peaks are very broad due to the adduction of sodium ions. After three steps of concentration-dilution using 10 kDa molecular weight cut off Millipore Amicron Ultra centrifuge filters, into 150 mM ammonium acetate, the sample was analysed again. The peaks were much narrower, and more separated, resulting in a higher mass resolution of the mass spectrum.

The resolving power of ToF mass analysers is constant across the *m/z* range [35]. The result of this is that the mass resolution is worsened as the *m/z* the analyte increases. Equation 2.8, shown previously, can be rearranged to calculate the mass resolution (Δm) from the ion *m/z* value (m) and the resolving power of the ToF (R), as seen in Equation 2.18.

$$\Delta m = \frac{m}{R} \tag{2.18}$$

Equations 2.19 and 2.20, show the calculation for mass resolution for a low and high *m/z* ion respectively, using the 10,000 R ToF used in the Waters Synapt G1 [26].

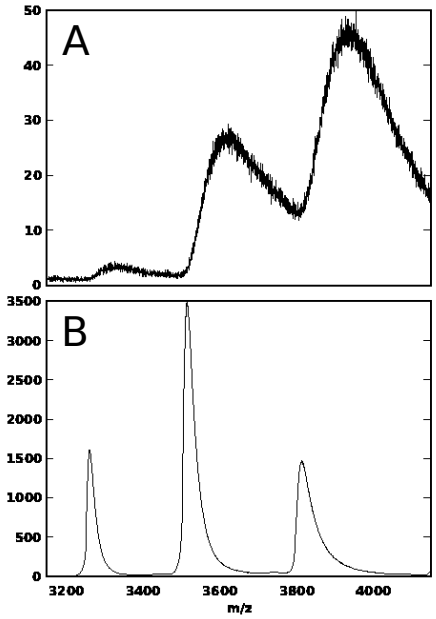

Figure 2.11: Mass spectra of a 45 kDa protein analysed under different buffer conditions; (A) analysed in 50 mM sodium phosphate buffer and 50 mM ammonium acetate. (B) same sample after 1,000 times dilution of original buffer into 150 mM ammonium acetate.

$$\frac{1000}{10000} = 0.1 \tag{2.19}$$

$$\frac{16000}{10000} = 1.6 \tag{2.20}$$

The 1000 m/z ion in Equation 2.19 has a mass resolution of 0.1, meaning that two ion peaks that are 0.1 m/z apart will be separated to half of their maximum height. Conversely when analysing an ion with m/z of 16,0000 (Equation 2.20) only peaks which are at least 1.6 m/z will be separated.

The resolution of linear ToFs is also limited due to variation in the initial energies of ions. When analysing ions of the same m/z value, more energetic ions reach the end of the ToF more quickly, as seen in Figure 2.12A. As the ToF mass analysers use time to measure m/z, this results in a spread of m/z being recorded for ions of a particular m/z value, as shown in Figure 2.12B.

In 1973, research out of the USSR [37] introduced the reflectron ToF which solved this issue. Ions are initially accelerated into the field-free region as

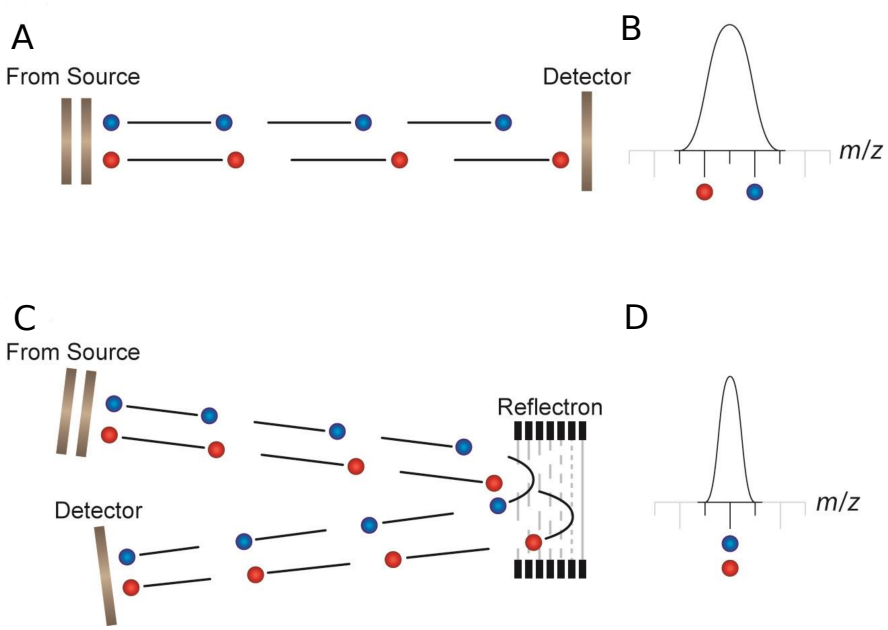

Figure 2.12: The difference between linear and reflectron time-of-flight (ToF) mass analysers. Molecules with the same m/z value will receive differing kinetic energy levels. Shown are two molecules with the same m/z value which receive less (blue) and more (less) kinetic energy. In a linear ToF (A) the ions reach the detector at different times and result in a broad mass spectrum peak (B). The reflectron (B) causes an equalisation of kinetic energy received by ions, leading to a narrower mass spectrum peak (D). Figure adapted from [2].

normal, but the ion beam is then reflected back by a second round of acceleration by electrode rings. The effect of this is that ions which received more energy at the initial acceleration stage go deeper into the reflectron region (Figure 2.12C). This equalises the total time of flight for ions with the same m/z, leading to the same m/z value being recorded, as seen in Figure 2.12D.

2.2.3 Tandem mass spectrometry

The use of multiple mass analysers allows for tandem mass spectrometry experiments, where a certain ion or m/z range is selected by the first mass analyser, and analysed by the second. Two common forms of instrumentation are triple quadrupole mass spectrometers, often shortened to triple-quad, and quadrupole time-of-flight mass spectrometers, abbreviated to Q-ToF and this report will focus on the latter.

Tandem mass spectrometry instrumentation is often combined with a collision cell, allowing for a technique called collision induced dissociation (CID), and was introduced in the 1960s by Jennings [38] and McLafferty [39].

The collision cell has an electric potential across it (often referred to as collision energy) and is filled with inert buffer gas. The electric field increases the kinetic energy of the ion, accelerating it through the collision cell. As the ion traverses the collision cell, the ions experience collisions with the buffer gas molecules, which causes the kinetic energy to be converted to internal energy through friction. This process is dependent on the pressure of the buffer gas, higher gas pressures results in more collisions, increasing ion heating. Similarly changing the voltage difference across the collision cell causes the ion to have more violent collisions with the gas molecules, thereby also increasing the internal energy of the ions.

The final factor that affects collisions is the m/z value. The force experienced by the ion is proportional to the charge of the ion, as per Newton's second law of motion, a lighter ion, with the same charge as a higher mass ion, will be accelerated faster through the collision cell. Thereby smaller ions will experience more violent collisions than larger ions.

The internal energy of ions can be increased to the extent that non-covalent interactions are broken. An example use of this, is increasing the collision energy in order to improve desolvation or for stripping adducted cations, which results in increased mass resolution.

Another use case for this technology is the sequencing of peptides. The quadrupole is used to select a particular ion, the collisions with the inert gas in the collision cell cause peptide bonds to fragment with different patterns, and the ToF is used to determine the mass of the fragments. Software is then used to determine the sequence of the original peptide [40].

CID can also be used to dissociate subunits of non-covalently bound multi-subunit proteins. A CID collision deposits energy onto the polypeptide chain, which then diffuses across it. The weakest interaction is eventually broken, in the case of peptide fragmentation the peptide bond fragments. Conversely in multiprotein complexes, the non-covalent interactions holding the complex

together are weaker and so are broken first. The data acquired for subunit dissociation demonstrates an asymmetric charge distribution, with the dissociated monomer having proportionally higher levels of protonation [41] which is not seen in solution thermal dissociation [42, 43].

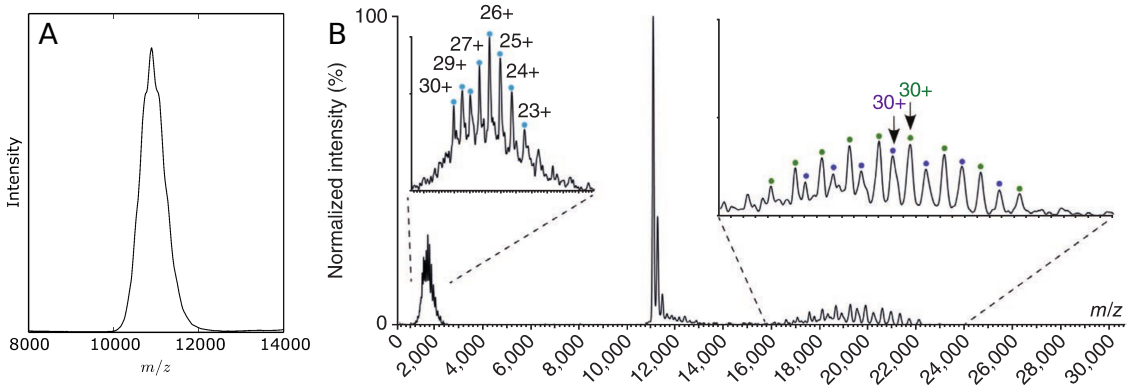

Figure 2.13: Collision induced dissociation (CID) used to analyse a heterogeneous sample with heavily overlapped peaks. Native condition mass spectrum of DegQ lysozyme complex (A). Mass spectrum of quadrupole isolated m/z region after CID (B - centre), with left inset showing the mass spectrum region of ejected apo DegQ monomer, and the right inset showing the charge reduced mass spectrum of 11-mer DegQ bound to 6 (green) and 5 (purple) lysozyme molecules. Figure adapted from [44].

Another use for the dissociation of non-covalent complexes, via CID, is when the charge state peaks of the complex in a mass spectrum are overlapping to an extent where they are indistinguishable. Using CID to dissociate a subunit from a complex results in uneven charge distribution, the reduced charge of the larger complex results in charge state peaks which have a greater difference in m/z value so that mass assignment is possible. Additionally, the ejected subunit is of low mass, and so experiences high energy collisions due to the high charge density, which usually results in narrow peaks which can be easily assigned. An example of this process was conducted by Malet *et al.* [44] and demonstrated in Figure 2.13.

2.2.4 Microchannel plate detector

The microchannel plate (MCP) detector is commonly used with ToF mass analysers as it has a very quick response time. The MCP detects the time of the collision, and the difference between when the ion enters the field-free region of the ToF and collides with the MCP is used to calculate the m/z ratio of an ion.

It is a circular plate with holes or channels bored into it at an angle. When an ion reaches the MCP it enters one of these channels and collides with the wall and the angle of the channels ensures this happens. This collision causes the channel to emit electrons, and as the electrons pass through the channel, additional collisions generate more electrons in a cascade. This amplifies the current generated by a single ion so that it is more easily detected. The fact that there are multiple channels means that several ion m/z values can be recorded simultaneously. A potential issue with this instrumentation, however is saturation. Preceding an ion impact event a channel will have a certain dead time before being able to register another ion. An additional feature is that as ions are separated by velocity, higher m/z ions hit the channels at lower velocity which results in a lower signal intensity being recorded.

Chapter 3

Ion mobility mass spectrometry

Ion mobility mass spectrometry (IM-MS) was first reported in 1962 by McDaniel *et al.*, which used a Nier 60° magnetic sector mass analyser [1], and in 1967 work at Bell Labs introduced the first ion mobility mass spectrometer which utilised a ToF for mass analysis [2, 3].

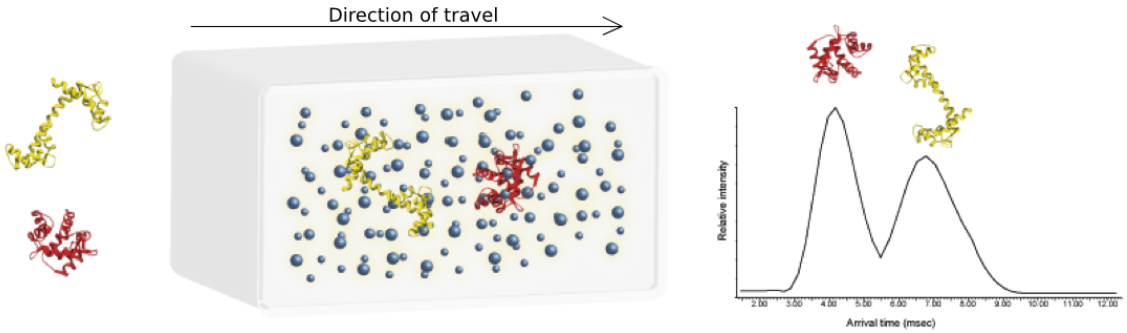

Figure 3.1: Schematic representation of ion mobility separation. The same protein with the same mass and charge is present in an open (yellow) and closed (red) conformation (left). As the ions traverse the ion mobility cell, the more closed conformation experiences less collisions, allowing it to traverse the cell faster (centre). The data are then acquired as an arrival time distribution (ATD), with time on the x axis and ion current on the y axis (right).

IM-MS adds another dimension of separation to mass spectrometry experiments, based upon the shape of ions. Ions are pulled through a chamber

filled with inert gas, typically helium [4] or nitrogen [5] by an electric field. As the ion traverses the chamber it experiences collisions with the gas, as seen in Figure 3.1 and the chance of a collision is dependent upon the cross section of the side of the ion facing the direction of travel. As the ion tumbles as it traverses the cell, the number of collisions experienced is dependent on the rotationally averaged collision cross section (CCS), which can be used to analyse the shape of an ion. It is measured as the time taken to traverse the ion mobility chamber and is dependent on the mass and charge of an ion, higher mass ions are accelerated less by the same force and the propulsive kinetic energy applied to the ion is proportional to its charge.

Ion mobility spectrometry (IMS) is an analytical technique in its own right [6] and is used in a wide range of applications including detection of explosives [7] and illicit drugs [8–10].

3.1 Drift tube ion mobility

Early IM-MS experiments for the analysis of proteins were carried out using drift tube ion mobility mass spectrometry (DTIM-MS) [11–14]. The path of ions through the drift tube chamber is linear and can be understood from physical principles. The ion separation occurs due to a difference in mobility (K), which is dependent on the constant electric field strength (E) and the averaged velocity for the ions to traverse the chamber (v_d) with the relationship shown in Equation 3.1.

$$K = \frac{v_d}{E} \tag{3.1}$$

The ion mobility can then be represented as reduced mobility (K_0), which is standardised to atmospheric pressure (760 Torr, P_0) and 0 °C (273.15 K, T_0). The relationship is described in Equation 3.2 where temperature is T and buffer gas pressure is P of the experiment.

$$K_0 = K \cdot \frac{P \cdot T_0}{T \cdot P_0} = K \cdot \frac{P \cdot 273.15}{T \cdot 760} \tag{3.2}$$

The CCS value (Ω) of an ion can then be calculated using the reduced ion mobility, reduced mass of the ion-gas pair (μ, Equation 3.3), buffer gas number density (N), Boltzmann's constant (k_B), temperature (T), the charge state of the ion (z) and the charge of an electron (e), using Equation 3.4.

$$\mu = \frac{m_{ion} \cdot m_{gas}}{m_{ion} + m_{gas}} \tag{3.3}$$

$$\Omega = \frac{3ze}{16N} \cdot \sqrt{\frac{2\pi}{\mu k_B T}} \cdot \frac{1}{K_0} \tag{3.4}$$

The instrumentation used for the analysis of proteins was often a quadrupole for mass analysis, followed by the drift tube [11]. In this case during ion mobility experiments it was only possible to monitor a single ion charge state, the quadrupole was used to select for a particular ion, and so only the ion mobility separation, as drift time was the data acquired.

As DTIM-MS records the drift time, it is not well suited to ESI experiments which generate a continuous flow of ions, whereas the pulsed ion flow of MALDI experiments work well. In order to overcome this issue, the ESI-DTIM-MS instrumentation was developed to utilise an ion trap interface before the drift cell, which stores ions and releases them at set intervals [15].

DTIM-MS instruments were developed *in house* generally by physical chemists, and required expertise in areas such as electronic engineering and software development. The result of this was that few labs were able to use the instrumentation.

3.2 Travelling wave ion mobility

In 2004 Giles *et. al* outlined a mass spectrometer that used a travelling wave to achieve ion mobility separation [16], and this system was later engineered into the first commerically available ion mobility mass spectrometer in 2006 [17]. The instrument included quadrupole and ToF mass analysers in addition to the travelling wave ion guide (Figure 3.2).

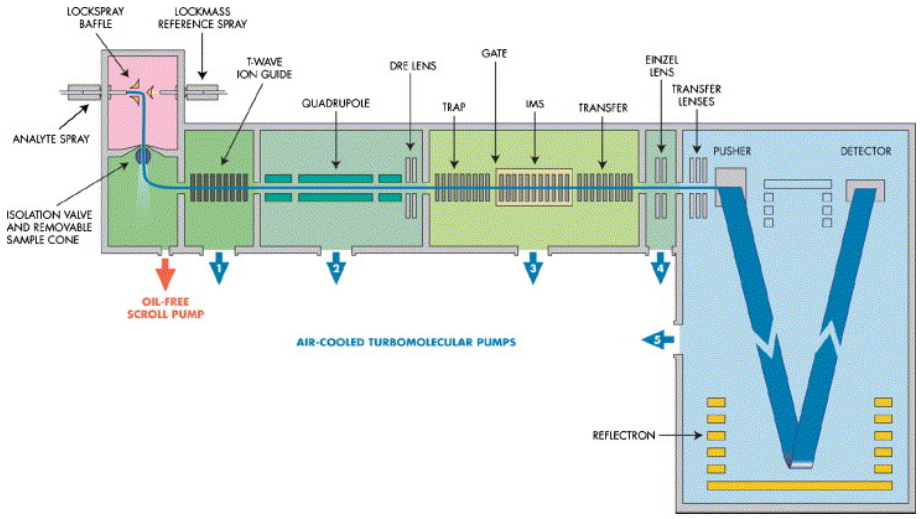

Figure 3.2: Schematic representation of the Waters Synapt travelling wave ion mobility mass spectrometer. Reproduced from [17].

The travelling wave system is known as a stacked ring ion guide (SRIG), and prevents the radial diffusion of ions whilst traversing the instrumentation. It is comprised of pairs of rings, which have opposite phases of RF voltage applied to each adjacent ring. The rings are arranged perpendicularly to the flow of ions so that the ions flow through the centre. The axial field created with increasing strength towards the ring, confines the ions within the ring orifice (as seen in Figure 3.3A).

The travelling wave instrumentation consists of three parts, the IM cell and the trap and transfer ion guides, which have 61, 31 and 31 pairs of ring electrodes respectively. The final pair of rings in the trap cell have a DC only potential applied to them, and this voltage is periodically modulated in order to trap or release ions into the IM cell [16]. The transfer cell is used to maintain ion mobility separation after the IM cell. The trap and transfer regions can be used as collision cells for CID by varying the voltage across

them.

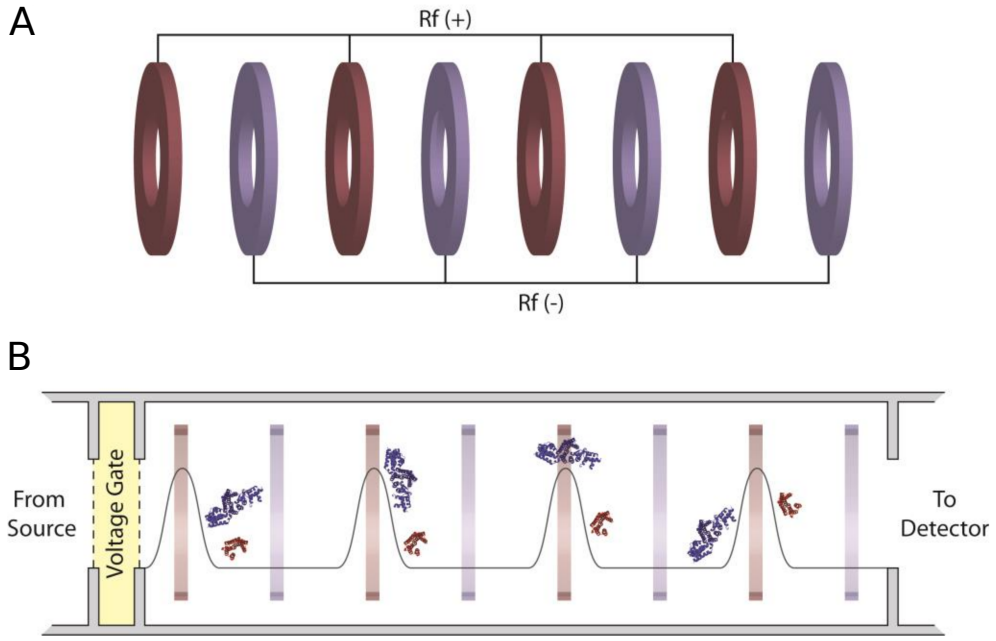

Figure 3.3: (A) Schematic representation of the stacked ring ion guide (SRIG), with opposite phase RF voltages being applied to adjacent rings. (B) Representation of how the travelling wave separates ions by size. After the release of the packet of ions at the trap ion guide voltage gate, ions are pushed through the ion mobility cell by the travelling wave, generated by the SRIG. Smaller ions are pushed straight through by the wave, whereas the movement of larger ions is retarded by more collisions with gas molecules causing the ion to go over the top of the wave. Figure adapted from [18].

The travelling wave is propagated by the SRIG and the RF voltage applied to it. The pairs of rings can combine to apply a positive or negative voltage, and so the amplitude of the travelling wave at a given time and position of the wave can be produced by a pair of ring electrodes. This allows the wave to work as a moving electric field pushing ions through the mass spectrometer [16].

The ion mobility cell can also be operated as purely an ion guide for standard mass spectrometry experiments. When in use for ion mobility experiments, it is filled with nitrogen gas, at up to 1 $m$bar pressure [17]. As the travelling wave moves through the IM cell smaller ions experience few collisions and so are pushed along by the wave, reducing the residence time in the cell. Conversely larger ions experience more collisions, which causes resis-

tance against the wave and so they can be pushed over the top of the wave, as demonstrated in Figure 3.3B. The system leads to impressive separation considering the IM cell is only 185 mm long [16].

A TWIM-MS experiment is made up of 200 separate mass spectra. This is achieved using the trapping functionality of the trap cell [19]. A packet of ions is released into the IM cell then the ion flow is stopped temporarily. The ions then traverse the travelling wave system, and a pusher injects ions into the ToF mass analyser 200 times for each packet of ions thereby acquiring mass spectra. The rate at which this happens is known as the pusher frequency (f) with the pusher interval normally in the micro second range. The total arrival time (t_t) measured in an IM experiment is given by Equation 3.5.

$$t_t = 200 \cdot f \tag{3.5}$$

This gives rise to multidimensional data as shown in Figure 3.4. Both the mass spectrum (Figure 3.4A) and arrival time data (Figure 3.4) are acquired simultaneously, with the intensity information for each data type being recorded as a grid, which can then be shown as a heatmap (Figure 3.4C).

Comparing drift tube and travelling wave IM-MS

The functionality of the travelling wave system to act as an ion guide leads to higher sensitivity than drift tube experiments. Drift tube separation instrumentation, however, tends to have higher mobility resolution. This is due to two factors, first the potential for very long drift tubes (potentially infinite due to the development of a circular drift cell [20]), increasing the length of a drift tube increases the difference in time separation, and so smaller changes are easier to detect. The second is the use of helium instead of nitrogen as mobility gas as well as the capability for higher gas pressures, smaller gas molecules act as a smaller probe features in the surface of the molecule, whereas larger gas molecules may not fit into crevices in the protein surface, thereby distorting the measured collision cross section.

Nitrogen is more polarisable ($1.740x10^{-24}$ cm^3) than helium ($0.205x10^{-24}$ cm^3). The polarisability of the drift gas used has been shown to change the

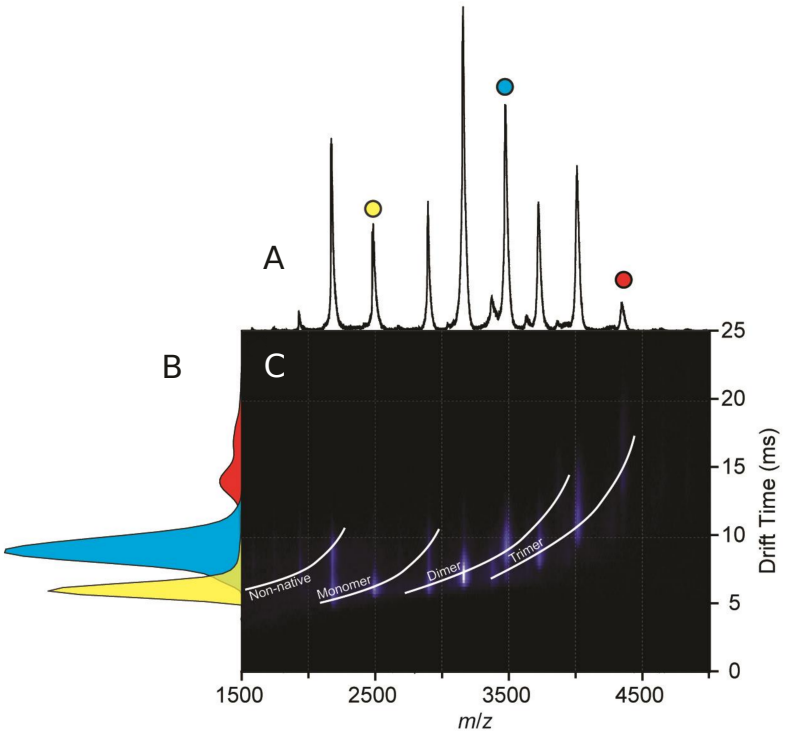

Figure 3.4: Example of the multidimensional data that can be acquired simultaneously using a Waters Synapt TWIM-MS instrument. Mass spectral data are shown, with the lowest charge state of the monomeric, dimeric and trimeric species are labelled with yellow, blue and red spheres respectively (A). Arrival time distributions for the lowest charge state of each oligomeric species are shown (B). The overall intensity heatmap of mass spectral and arrival time data also shown (C). Figure adapted from [18].

reported CCS value of analytes in IM-MS experiments. Figure 3.5 shows the effect with drift gases of varying levels of polarisability on glycine peptides. It can be seen that as the size of the peptide increases, the effect is lessened, with the largest peptide analysed being 360 Da [21]. Further work by Ruotolo and co-workers investigated this effect on the tryptically digested proteins, and found there to be no appreciable effect on peptides analysed in the 500-3500 m/z region [22]. None of the samples investigated here are below 500 m/z so this effect should not change the results.

Drift tube experiments typically use mass analysis, in the form of a quadrupole, to select an ion, and then record an arrival time distribution for the ion. A benefit of TWIM-MS is the ability to acquire mass spectra, using the ToF, at the same time as recording arrival time. This gives rise to

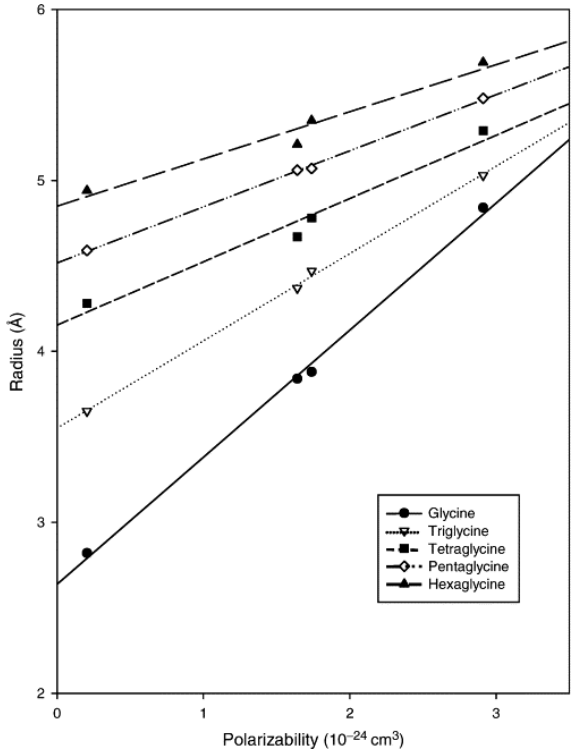

Figure 3.5: Effect on recorded radius of glycine oligomers using ion mobility mass spectrometry. The data points in the series indicate the polarisability of the buffer gas used, from low to high; helium, argon, nitrogen and carbon dioxide. Figure reproduced from [21].

the multidimensional data shown in Figure 3.4, which is not possible using most drift cell instruments.

The path of an ion travelling through a drift tube can be understood from physical principles, and mobility and CCS have a linear relationship. This allows for direct calculation of CCS values from these data. The path of an ion through a TWIM-MS instrument is complicated due to the effect of the travelling wave, this is compounded by the non-linear relationship between arrival time and ion CCS [23], which is evidenced by the curved arrival times of charge state arrival time distributions in Figure 3.4C. The result of this is that it has not been possible to directly calculate the CCS values of ions from TWIM-MS data [24].

TWIM-MS calibration

Travelling wave ion mobility mass spectrometry arrival time data can be calibrated using ions of known CCS by several methods [5, 25, 26].

The CCS values used for calibration are generally taken from experimental data, acquired using drift tube mass spectrometers. Early on most TWIM-MS calibrations were conducted using peptides or proteins in denaturing buffer [5, 19, 27], with particular interest in small proteins such as denatured myoglobin [25]. The use of small denatured proteins resulted in a mass spectrum with many charge states covering a large t_d range, and these two effects helped create good calibration curves. Clemmer and co-workers developed the Clemmer online database*, which compiled several publications worth of experimental CCS values for denatured proteins [12, 28, 29].

In 2012 Salbo *et al.* discovered that the calculated CCS values for native proteins was more accurate when using native proteins as calibrants [30]. The research was facilitated by a publication by Bush *et al.* which gave the CCS values for several native and denatured proteins [31]. The study used a Waters Synapt instrument which had the travelling wave portion replaced by a drift tube, this allowed several contributing factors to CCS determination to be removed when comparing it to the same instrument with a travelling wave installed. Another interesting feature of the dataset is that CCS values were recorded for the calibrants using both helium and nitrogen (as used in travelling wave) as the mobility gas. Several native and denatured proteins were analysed and the collection in kept up-to-date on the Bush CCS Database†.

The different published methods for calibrating TWIM-MS data use differing equations to calculate CCS however they result in the same answer. An explanation of the calibration process described by the Scrivens group [5] is described here.

The time taken for the analyte to traverse the mass spectrometer after the trap ion guide, through the IM separation cell, transfer ion guide and ToF is given as a scan number (n). Conversion of this value to milliseconds is achieved using Equation 3.6 using the pusher frequency (f).

$$t_d = n \cdot f \tag{3.6}$$

*http://www.indiana.edu/~clemmer/Research/Cross%20Section%20Database/cs_database.php
†http://depts.washington.edu/bushlab/ccsdatabase/

As the arrival time value has contributions which are not involved in IM separation, these factors have to be corrected for. These factors are split into those which are dependent ($t_{dependent}$) and independent ($t_{independent}$) of the m/z of the ion and so the correction is made using Equation 3.7.

$$t'_d = t_d - t_{independent} - t_{dependent} \tag{3.7}$$

The m/z independent factors are contributed to by the time spent traversing the ion mobility and transfer cells. The size of the contribution is related to the wave velocity of the travelling wave (V_w), and the length of each of the cells. The time taken to traverse a pair of stacked ring electrodes is 0.010 ms at a wave velocity of 300 m/s. The ion mobility cell has 61 electrode pairs and the transfer has 31, the relationship is described in Equation 3.8.

$$t_{independent} = (61 + 31)\left(0.01 \cdot \frac{300}{V_w}\right) \tag{3.8}$$

The m/z dependent contribution to arrival time is the time taken for the analyte ion to reach the detector after exiting the transfer cell. An ion with an m/z value of 1,000 takes 0.044 ms to get from the T-wave to ToF and a further 0.041 ms to reach the detector. The time is proportional to the square root of the m/z value and the m/z dependent portion of the arrival time can be calculated with Equation 3.9.

$$t_{dependent} = \left(\sqrt{\frac{m/z}{1000}} \cdot (0.044 + 0.041)\right) \tag{3.9}$$

The experimental CCS values of the calibrants have to be corrected for reduced mass (μ) using the mass of the ion (m_{ion}) and mass of the gas (m_{gas}), and the charge state (z) (Equations 3.3 and 3.10).

$$\Omega' = \Omega/z \cdot \sqrt{\frac{1}{\mu}} \tag{3.10}$$

A power fit is then applied to the corrected CCS value (Ω') against corrected arrival time calibration curve using Equation 3.11) to determine the two fit coefficients A and B. An example calibration curve for denatured myoglobin is shown in Figure 3.6.

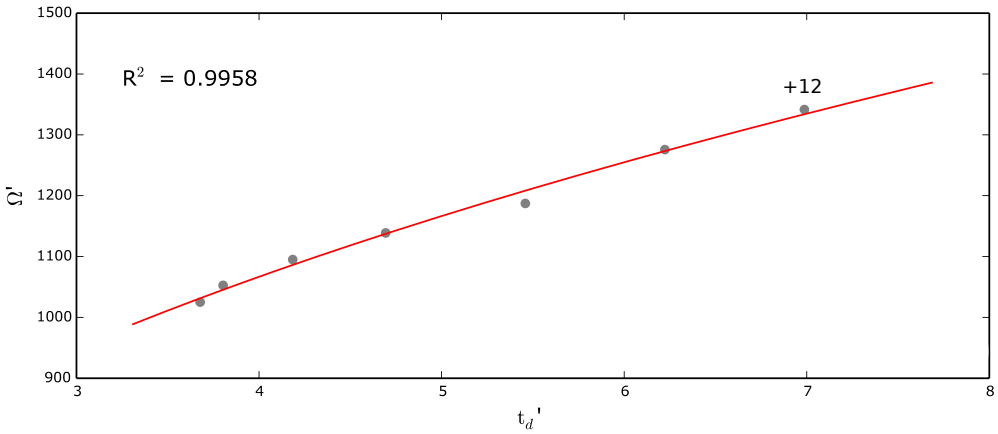

Figure 3.6: A TWIM-MS calibration curve using denatured myoglobin as the calibrant. The plot shows corrected arrival time (t_d') against corrected CCS (Ω'), and the curve fits the data with an R^2 value of 0.9958.

$$\Omega' = A \cdot t_d'^{B} \tag{3.11}$$

The previous equations can be combined to give Equation 3.12. Using the power fit coefficients, the CCS value of an analyte can be calculated and the value is measured in squared Angstroms (Å^2).

$$\Omega = A \cdot t_d'^{B} \cdot z \cdot \sqrt{\frac{1}{\mu}} \tag{3.12}$$

In drift tube experiments, careful monitoring of gas pressure, composition and temperature is required to acquire reliable data. However as TWIM-MS are compared to calibrant proteins which are acquired under the same conditions, usually during the same experimental session, these factors can be left out of the calibration procedure [30]. The error in CCS determination using TWIM-MS was additionally shown to be under 5 % by Bush *et al.* [31] and work by Leary *et al.* has shown that the CCS values calculated have good

reproducibility when being analysed at separate laboratories [32].

3.3 Collision cross section calculations

As ion mobility mass spectrometry is able to report a physical characteristic of the shape of proteins, in the form of CCS, it was necessary to compare this to other structural biology techniques for analysis. The Protein Databank (PDB)[‡] is a repository of protein structures, in the form of nuclear magnetic resonance (NMR), X-ray crystallography and model structures. X-ray crystallography and NMR data are the most detailed information available for the structure of proteins. The data are in the form of three dimensional coordinates for each atom in the structure. In order to be able to compare the PDB structures to IM-MS CCS data, it is necessary to calculate the CCS of proteins represented as coordinates. There have been several proposed approaches to calculating this value and some of the most widely used methods are discussed here.

Projection approximation

The fastest method of calculating the CCS of a PDB structure is using the projection approximation algorithm (PA) [33]. In the algorithm, the atoms of a protein are treated as opaque spheres, using Van der Walls radii as the radii of the spheres. A 2D virtual grid with x and y axes is then created and the model is projected onto it by collapsing the z axis, creating a 'shadow' [34]. Each square on the grid is then checked for whether it is occupied by an atom. The number of squares containing atoms is then counted and multiplied to give the cross sectional area of the projection. The protein is then rotated and this process is repeated. The average CCS value of each rotation is calculated and when the average converges within a certain tolerance level (e.g. 1 $Å^2$), then the algorithm exits. An alternative method for calculating the area is using Monte Carlo integration [35], in this case the z axis is still collapsed, but now x and y coordinates are randomly chosen and checked whether they fall in or outside of the area of the protein. To calculate the area, the percentage

[‡]http://www.rcsb.org/pdb/

of hits is multiplied by the area of the x and y axes grid, points keep being generated and tested until the area of the single projection converges.

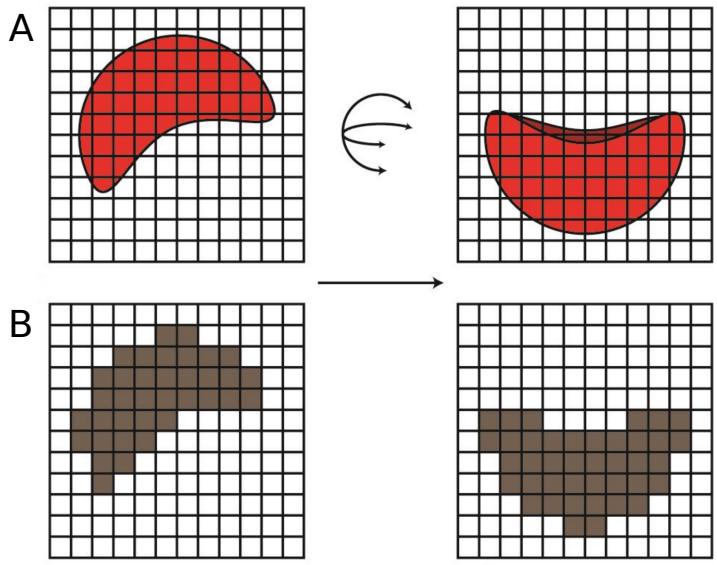

Figure 3.7: An illustration of the projection approximation (PA) algorithm. The ion, shown in red (A), is projected onto a grid (B), the number of grid squares covered by the ion is counted and multiplied by the area of the square to get the area of the projection. The ion is then rotated, and the process repeated (right), until the average area of projections converges. Figure adapted from [18].

The PA is known to underestimate CCS values of protein structures. This is because of the lack of gas molecules in the algorithm model. Long range electrostatic interactions are ignored, as well as the potential for multiple collisions for a single gas molecule, which is especially important in the case of convex and complex protein shapes. This together suggests theoretically that the algorithm underestimates the retardation caused by collisions with the buffer gas, and this has also been shown to be true empirically [19, 32].

The PA algorithm is included in several software packages including Sigma[§], MOBCAL[¶], and recently a fast PA algorithm has been developed by Erik Marklund from the Benesch group and though is as yet unpublished, is available to download[‖].

[§]http://bowers.chem.ucsb.edu/theory_analysis/cross-sections/sigma.shtml
[¶]http://www.indiana.edu/~nano/software.html
[‖]http://impact.chem.ox.ac.uk/

Trajectory method

The trajectory method (TM) is the most accurate method for calculating the CCS of a PDB structure [19, 35]. In this model gas molecules are fired at the protein, and all interactions are taken into account, including Lennard-Jones potentials and long range electrostatic interactions [33], and as with the PA, the collisional area is calculated for many rotations to give the orientally averaged CCS value [36]. The benefit of accuracy comes at the cost of computation time, with calculations for large proteins being unfeasible even on modern hardware, with some proteins taking months to analyse [35].

Exact hard sphere scattering

The exact hard sphere scattering algorithm (EHSS), was introduced as a middle ground between the PA and TM. In this algorithm the atoms are treated as hard spheres and no long range interactions are considered. The algorithm fires gas molecules at the protein and determines if it passes straight past or makes a collision. The angle of deflection is calculated, and the path of gas molecule is followed, the collisional area is further increased if the gas molecule makes an additional collision. This method is known to overestimate the CCS value in comparison to the TM [32], though it is thought that the accuracy is reduced by only a few percent [19, 37].

The simplifications of the model used in the EHSS result in a vastly reduced computation time in comparison to the TM. Even though it does take longer than the PA, CCS values for proteins take in the order of minutes to compute.

Projection superposition approximation

The projection superposition approximation algorithm (PSA) [35, 38, 39], is a recently adapted version of the PA. The PSA uses a shape factor to estimate the extent of the convex nature of a protein, and calculates collision probabilities rather than using hard spheres. The work has shown the algorithm to be considerably faster than the TM and calculates CCS values which are closest to the TM of the algorithms presented here [35].

3.4 Gas-phase protein conformation

The success in ion mobility instrumentation raised the question of whether proteins retained their native structure when analysed *in vacuo*. Early studies by the Clemmer and Jarrold groups studied this question conducting electrospray ionisation drift tube ion mobility experiments on cytochrome *c* and comparing the results to the calculated CCS values of X-ray crystallography structures. In 1995 it was found that the IM analysis yielded a larger CCS value than expected of natively folded cytochrome *c*, this however was largely due to the use of 50:50 methanol to water as the solvent [11]. Work in 1997 resulted in very close correlation between experimental and theoretical CCS values, even with the use of acetonitrile in the solvent [12, 40]. The work found that the lowest charge states of ions have the closest CCS value to the theoretical, as additional charges can cause unfolding due to coulombic repulsion. It was also demonstrated that altering the velocity of ions entering the drift tube, and so the energy gained from collisions with gas molecules (known as injection energy), would cause unfolding and increases in CCS, thereby suggesting that these voltages should be kept to a minimum in order to maintain protein structure.

After the introduction of nanoelectrospray ionisation, aqueous ammonium acetate could be used as solvent, making it easier to achieve native-like CCS values by ion mobility and an example of this was in 2004 when Bernstein *et al.* analysed α synuclein [4]. The introduction of TWIM-MS and work from the Scrivens laboratory brought further validation of the possibility of analysing proteins in native-like conformations [19, 41]. It was later shown that proteins of larger and more complex shapes could be maintained. A demonstration of this was shown in work done on *trp* RNA binding attenuation protein (TRAP), the subunits of which form an 11-mer ring structure ($\sim$80 kDa) which was maintained in the gas-phase [42]. These advances in understanding allowed the Bowers group to elucidate the repeating structure in Aβ42 fibrils which are responsible for Alzeihmer's disease by comparing experimental ion mobility data to molecular models of potential structures [43].

3.5 TWIM-MS experiments

An introduction to the methodology used to carry out TWIM-MS experiments on a Waters Synapt G1 follows. First an experiment aiming to compare with protein structure models is described followed by a collision induced unfolding experiment.

Native structure TWIM-MS

Analysis of proteins using TWIM-MS experiments can achieve very similar conformations to that of X-ray crystallography experiments.

In order to maintain a native-like protein conformation, it is necessary to not expose the analyte to excessively violent collisions which causes collisional heating and can result in the breakage of hydrogen bonds and the unfolding of proteins or the dissociation of subunits from protein complexes. This can be controlled by the user by changing the various voltages in the mass spectrometer. The voltage of the capillary voltage and the cone voltages at the entrance to the mass spectrometer maintain the potential difference that allows electrospray to occur, must be above a certain value (dependent on the sample) in order to maintain the ion flow. Once electrospray has been achieved, these voltages should be optimised to be as low as possible whilst maintaining the electrospray. The trap voltage and transfer voltages should also be minimised. The acceleration provided by the trap voltage determines the initial velocity of the ions as they enter the ion mobility cell and so should typically be higher than the transfer voltage.

The nitrogen buffer gas pressure is also important in the trap and IM cells. The IM buffer gas provides the separation of ions based on the CCS, and so increasing the buffer gas pressure, increases the length of time taken to traverse the IM cell and typically improves resolution as a result. The high pressure of the IM cell ($\tilde{}5 \times 10^{-1} m$bar)would create a very large pressure difference between the trap and IM cells, which would disrupt ion flow. To combat this, during TWIM-MS experiments the gas pressure in the trap cell is also increased.

The travelling wave settings should be optimised to achieve the best separation. This is primarily achieved by changing the wave height (voltage) and wave velocity ($m \cdot s^{-1}$) of the IM cell. The wave height changes the probability of an ion rolling over the top of the T-wave or being carried along with it. It is important to optimise this for maximum resolution, this should be done with care as higher voltage wave heights can cause protein unfolding. The wave velocity increases the number of individual waves an ion experiences as it traverses the IM cell, which also improves the data. This must be optimised whilst observing the resulting ATD. Though higher wave velocities improve the effectiveness of the T-wave, it also results in the ions traversing the IM cell faster which results in a narrower ATD. As there are only 200 scans per ATD, the ATD of the ion should span as many of the scans as possible to acquire high resolution data. The optimisation process for the IM cell settings then has to include the buffer gas pressure, T-wave wave velocity and height at the same time to acquire the highest resolution data.

The resulting arrival time data can then be calibrated as shown in Section 3.2 and be compared to models.

These types of experiment are becoming increasingly prevalent to study protein complexes that are not amenable to analysis by X-ray crystallography or nuclear magnetic resonance (NMR). X-ray crystallography and NMR analysis becomes difficult for very large protein complexes [44], and also the structure of highly labile regions often is not determined, or have to be removed in order to form crystals [45]. Both types of experiment have difficulties with analysing heterogenous samples and require high protein concentrations [46].

IM-MS is able to analyse heterogenous samples as the components in the analyte solution are usually separated by mass, allowing the ion mobility analysis of a particular analyte in isolation. Very low sample concentrations are required (˜10 mM), and protein complexes of masses of over 400 kDa can be analysed [45].

The result of this is that ion mobility mass spectrometry may become a vital tool in structural proteomics, that has the aim of determining the structure

of every protein in the proteome. By IM-MSthe CCS value of individual subunits, sets of subunits and the total structure can be analysed. This can be combined with *in silico* molecular modelling in order to determine high resolution final structures, and would help with analysing protein complexes which cannot be analysed by other structural techniques [44].

Collision induced unfolding

Collision induced unfolding (CIU) experiments, involve increasing the collision energy in the trap cell in order to unfold the protein. When carrying out these experiments it is first necessary to optimise the settings for maintaining a native conformation (as described above) to have as the starting point of the experiment. The voltage should be increased until the signal of the ion peak being analysed disappears. It should then be the checked that the travelling wave parameters used are appropriate by examining an ATD of the ion. It may be the case that roll over occurs, which manifests as baseline lift. If this is observed then the travelling wave settings should be changed, typically this would involve reducing the wave height and/or increasing the wave velocity.

With the settings of the ion mobility mass spectrometer setup, the data can be acquired. The trap voltage should be increased at a user defined increment, acquiring each voltage step. It is important that the other settings are not changed during the CIU experiment. Altering the buffer gas pressure or travelling wave settings will cause changes in the resulting ATDs and changing other collision energies will alter the degree of unfolding of the protein. The experiment produces a set of ATDs for the ion peak which can be analysed further.

One method of representing these data is known as CIU fingerprinting and an example is shown in Figure 3.8. This representation consists of stacking ATDs at increasing collision energy along the x axis, with the CCS or arrival time (referred to in the figure as drift time) along the y axis and the intensity along the z axis and represented by colours. In this experiment, researchers investigated the effect of salt adducts on the gas phase stability of the tetrameric protein complexes avidin and concanavalin A (ConA).

The portion of the experiment shown is the comparison between the pro-

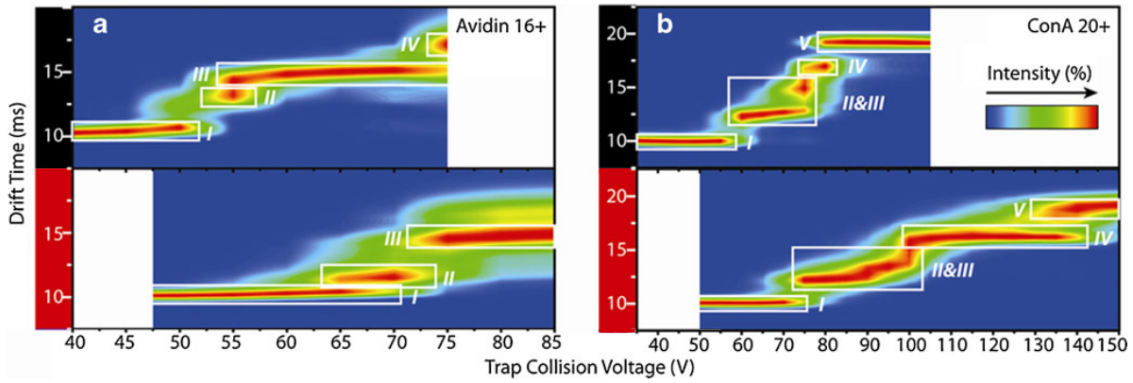

Figure 3.8: CIU finger print analysis of avidin (a) and concanavalin A (b), in 100 mM ammonium acetate (above) and 100 mM ammonium acetate with 2 mM magnesium acetate (below). Figure adapted from [47].

teins in ammonium acetate solution as well as with added magnesium acetate. The use of magnesium acetate meant that the increased adduction was only due to adduction of cations, as acetate ions are volatile and would evaporate. The increased resistance to unfolding, is indicated by reduced increase drift time as the collision energy is increased. This feature is shown by both proteins in the presence of magnesium acetate, which shows that magnesium ions had stabilised the conformation of the protein complexes [47].

Collision induced unfolding experiments will be investigated further in Chapter 5.

References

[1] McDaniel, E. W., Martin, D. W., and Barnes, W. S. (1962). "Drift tube-mass spectrometer for studies of low-energy ion-molecule reactions". *Review of Scientific Instruments* 33.1, pp. 2–7.

[2] McAfee, K. B., Sipler, D., and Edelson, D. (1967). "Mobilities and Reactions of Ions in Argon". *Physical Review* 160.1, pp. 130–135.

[3] Edelson, D., Morrison, J. A., McKnight, L. G., and Sipler, D. P. (1967). "Interpretation of Ion-Mobility Experiments in Reacting Systems". *Physical Review* 164.1, pp. 71–75.

[4] Bernstein, S. L., Liu, D., Wyttenbach, T., Bowers, M. T., Lee, J. C., Gray, H. B., and Winkler, J. R. (2004). "α-synuclein: Stable compact and extended monomeric structures and pH dependence of dimer formation". *Journal of the American Society for Mass Spectrometry* 15.10, pp. 1435–1443.

[5] Thalassinos, K., Grabenauer, M., Slade, S. E., Hilton, G. R., Bowers, M. T., and Scrivens, J. H. (2009). "Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry". *Analytical Chemistry* 81.1, pp. 248–254.

[6] Cohen, M. J. and Karasek, F. W. (1970). "Plasma Chromatography™—A New Dimension for Gas Chromatography and Mass Spectrometry". *Journal of Chromatographic Science* 8.6, pp. 330–337.

[7] Ewing, R. G., Atkinson, D. A., Eiceman, G. A., and Ewing, G. J. (2001). "A critical review of ion mobility spectrometry for the detection of explosives and explosive related compounds". *Talanta* 54.3, pp. 515–529.

[8] Keller, T., Miki, A., Regenscheit, P., Dirnhofer, R., Schneider, A., and Tsuchihashi, H. (1998). "Detection of designer drugs in human hair by ion mobility spectrometry (IMS)". *Forensic Science International* 94.1–2, pp. 55–63.

[9] Fytche, L., Hupé, M, Kovar, J., and Pilon, P (1992). "Ion mobility spectrometry of drugs of abuse in customs scenarios: concentration and temperature study." *Journal of Forensic Sciences* 37.6, pp. 1550–1566.

[10] March, R. E. and Todd, J. F. J. (2009). *Practical Aspects of Trapped Ion Mass Spectrometry: Applications of Ion Trapping Devices*. CRC Press.

[11] Clemmer, D. E., Hudgins, R. R., and Jarrold, M. F. (1995). "Naked Protein Conformations: Cytochrome c in the Gas Phase". *Journal of the American Chemical Society* 117.40, pp. 10141–10142.

[12] Shelimov, K. B., Clemmer, D. E., Hudgins, R. R., and Jarrold, M. F. (1997). "Protein structure in vacuo: Gas-phase conformations of BPTI and cytochrome *c*". *Journal of the American Chemical Society* 119.9, pp. 2240–2248.

[13] Wyttenbach, T., Helden, G. von, and Bowers, M. T. (1996). "Gas-Phase Conformation of Biological Molecules: Bradykinin". *Journal of the American Chemical Society* 118.35, pp. 8355–8364.

[14] Jarrold, M. F. (2000). "Peptides and proteins in the vapor phase". *Annual Review of Physical Chemistry* 51.1, pp. 179–207.

[15] Hoaglund, C. S., Valentine, S. J., and Clemmer, D. E. (1997). "An Ion Trap Interface for ESIIon Mobility Experiments". *Analytical Chemistry* 69.20, pp. 4156–4161.

[16] Giles, K., Pringle, S. D., Worthington, K. R., Little, D., Wildgoose, J. L., and Bateman, R. H. (2004). "Applications of a travelling wave-based radio-frequency-only stacked ring ion guide". *Rapid Communications in Mass Spectrometry* 18.20, pp. 2401–2414.

[17] Pringle, S. D., Giles, K., Wildgoose, J. L., Williams, J. P., Slade, S. E., Thalassinos, K., Bateman, R. H., Bowers, M. T., and Scrivens, J. H. (2007). "An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument". *International Journal of Mass Spectrometry* 261.1, pp. 1–12.

[18] Kerr, R. A. (2013). "Ion mobility & mass spectrometric studies of macromolecules required for organism viability". Doctoral thesis. University College London.

[19] Scarff, C. A., Thalassinos, K., Hilton, G. R., and Scrivens, J. H. (2008). "Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements". *Rapid Communications in Mass Spectrometry* 22.20, pp. 3297–3304.

[20] Bohrer, B. C., Merenbloom, S. I., Koeniger, S. L., Hilderbrand, A. E., and Clemmer, D. E. (2008). "Biomolecule Analysis by Ion Mobility Spectrometry". *Annual Review of Analytical Chemistry* 1, pp. 293–327.

[21] Beegle, L. W., Kanik, I., Matz, L., and Hill Jr., H. H. (2002). "Effects of drift-gas polarizability on glycine peptides in ion mobility spectrometry". *International Journal of Mass Spectrometry* 216.3, pp. 257–268.

[22] Ruotolo, B. T., McLean, J. A., Gillig, K. J., and Russell, D. H. (2004). "Peak capacity of ion mobility mass spectrometry: the utility of varying drift gas polarizability for the separation of tryptic peptides". *Journal of Mass Spectrometry* 39.4, pp. 361–367.

[23] Shvartsburg, A. A. and Smith, R. D. (2008). "Fundamentals of traveling wave ion mobility spectrometry". *Analytical Chemistry* 80.24, pp. 9689–9699.

[24] Smith, D. P., Knapman, T. W., Campuzano, I., Malham, R. W., Berryman, J. T., Radforda, S. E., and Ashcroft, A. E. (2009). "Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies". *European Journal of Mass Spectrometry* 12.13, p. 13.

[25] Ruotolo, B. T., Benesch, J. L., Sandercock, A. M., Hyung, S.-J., and Robinson, C. V. (2008). "Ion mobility–mass spectrometry analysis of large protein complexes". *Nature Protocols* 3.7, pp. 1139–1152.

[26] Williams, J. P. and Scrivens, J. H. (2008). "Coupling desorption electrospray ionisation and neutral desorption/extractive electrospray ionisation with a travelling-wave based ion mobility mass spectrometer for the analysis of drugs". *Rapid Communications in Mass Spectrometry* 22.2, pp. 187–196.

[27] Hilton, G. R., Thalassinos, K., Grabenauer, M., Sanghera, N., Slade, S. E., Wyttenbach, T., Robinson, P. J., Pinheiro, T. J., Bowers, M. T., and Scrivens, J. H. (2010). "Structural analysis of prion proteins by means of drift cell and traveling wave ion mobility mass spectrometry". *Journal of the American Society for Mass Spectrometry* 21.5, pp. 845–854.

[28] Valentine, S. J., Counterman, A. E., and Clemmer, D. E. (1997). "Conformer-Dependent Proton-Transfer Reactions of Ubiquitin Ions". *Journal of the American Society for Mass Spectrometry* 8.9, pp. 954–961.

[29] Valentine, S. J., Anderson, J. G., Ellington, A. D., and Clemmer, D. E. (1997). "Disulfide-intact and -reduced lysozyme in the gas phase: conformations and pathways of folding and unfolding". *The Journal of Physical Chemistry B* 101.19, pp. 3891–3900.

[30] Salbo, R., Bush, M. F., Naver, H., Campuzano, I., Robinson, C. V., Pettersson, I., Jørgensen, T. J. D., and Haselmann, K. F. (2012). "Travelling-wave ion mobility mass spectrometry of protein complexes: accurate calibrated collision cross-sections of human insulin oligomers". *Rapid Communications in Mass Spectrometry* 26.10, pp. 1181–1193.

[31] Bush, M. F., Hall, Z., Giles, K., Hoyes, J., Robinson, C. V., and Ruotolo, B. T. (2010). "Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology". *Analytical Chemistry* 82.22, pp. 9557–9565.

[32] Leary, J. A., Schenauer, M. R., Stefanescu, R., Andaya, A., Ruotolo, B. T., Robinson, C. V., Thalassinos, K., Scrivens, J. H., Sokabe, M., and Hershey, J. W. (2009). "Methodology for measuring conformation of solvent-disrupted protein subunits using T-WAVE ion mobility MS:

an investigation into eukaryotic initiation factors". *Journal of the American Society for Mass Spectrometry* 20.9, pp. 1699–1706.

[33] Mesleh, M. F., Hunter, J. M., Shvartsburg, A. A., Schatz, G. C., and Jarrold, M. F. (1996). "Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential". *The Journal of Physical Chemistry* 100.40, pp. 16082–16086.

[34] Scarff, C. A., Patel, V. J., Thalassinos, K., and Scrivens, J. H. (2009). "Probing Hemoglobin Structure by Means of Traveling-Wave Ion Mobility Mass Spectrometry". *Journal of the American Society for Mass Spectrometry* 20.4, pp. 625–631.

[35] Bleiholder, C., Wyttenbach, T., and Bowers, M. T. (2011). "A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (I). Method". *International Journal of Mass Spectrometry* 308.1, pp. 1–10.

[36] Campuzano, I., Bush, M. F., Robinson, C. V., Beaumont, C., Richardson, K., Kim, H., and Kim, H. I. (2012). "Structural Characterization of Drug-like Compounds by Ion Mobility Mass Spectrometry: Comparison of Theoretical and Experimentally Derived Nitrogen Collision Cross Sections". *Analytical Chemistry* 84.2, pp. 1026–1033.

[37] Jarrold, M. F. (1999). "Unfolding, refolding, and hydration of proteins in the gas phase". *Accounts of Chemical Research* 32.4, pp. 360–367.

[38] Anderson, S. E., Bleiholder, C., Brocker, E. R., Stang, P. J., and Bowers, M. T. (2012). "A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (III): Application to supramolecular coordination-driven assemblies with complex shapes". *International Journal of Mass Spectrometry* 330–332, pp. 78–84.

[39] Bleiholder, C., Contreras, S., Do, T. D., and Bowers, M. T. (2013). "A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (II). Model parameteriza-

tion and definition of empirical shape factors for proteins". *International Journal of Mass Spectrometry* 345–347, pp. 89–96.

[40] Clemmer, D. E. and Jarrold, M. F. (1997). "Ion mobility measurements and their applications to clusters and biomolecules". *Journal of Mass Spectrometry* 32.6, pp. 577–592.

[41] Thalassinos, K., Slade, S. E., Jennings, K. R., Scrivens, J. H., Giles, K., Wildgoose, J., Hoyes, J., Bateman, R. H., and Bowers, M. T. (2004). "Ion mobility mass spectrometry of proteins in a modified commercial mass spectrometer". *International Journal of Mass Spectrometry* 236.1, pp. 55–63.

[42] Ruotolo, B. T., Giles, K., Campuzano, I., Sandercock, A. M., Bateman, R. H., and Robinson, C. V. (2005). "Evidence for macromolecular protein rings in the absence of bulk water". *Science* 310.5754, pp. 1658–1661.

[43] Bernstein, S. L., Dupuis, N. F., Lazo, N. D., Wyttenbach, T., Condron, M. M., Bitan, G., Teplow, D. B., Shea, J.-E., Ruotolo, B. T., Robinson, C. V., and Bowers, M. T. (2009). "Amyloid-β protein oligomerization and the importance of tetramers and dodecamers in the aetiology of Alzheimer's disease". *Nature Chemistry* 1.4, pp. 326–331.

[44] Zhong, Y., Hyung, S.-J., and Ruotolo, B. T. (2012). "Ion mobility–mass spectrometry for structural proteomics". *Expert Review of Proteomics* 9.1, pp. 47–58.

[45] Politis, A., Park, A. Y., Hyung, S.-J., Barsky, D., Ruotolo, B. T., and Robinson, C. V. (2010). "Integrating ion mobility mass spectrometry with molecular modelling to determine the architecture of multiprotein complexes". *PLoS One* 5.8, e12080.

[46] Zhong, Y., Hyung, S.-J., and Ruotolo, B. T. (2011). "Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions". *Analyst* 136.17, pp. 3534–3541.

[47] Han, L. and Ruotolo, B. T. (2013). "Traveling-wave ion mobility-mass spectrometry reveals additional mechanistic details in the stabilization of

protein complex ions through tuned salt additives". *International Journal for Ion Mobility Spectrometry* 16.1, pp. 41–50.

Chapter 4

Amphitrite

Ion mobility mass spectrometry can provide multidimensional data for protein analytes. This chapter introduces new methods of data analysis and representation; these use heat maps to incorporate the extra dimension provided by IM-MS into the data representation, and improve the reproducibility and accuracy of data extraction and manipulation.

4.1 Introduction

4.1.1 Deconvolution of ESI mass spectra

Electrospray ionisation experiments produce spectra with several individual peaks owing to differing numbers of charges. This is different to other methods such as matrix assisted laser desorption ionisation (MALDI) experiments which typically produce a single peak corresponding to the singly charged component. These extra peaks increase mass accuracy as the value taken is an average, however the more crowded spectra, when multiple molecular species are present, can make data analysis considerably more complicated.

The experiments described here were carried out in positive ion mode, meaning that the multiple peaks were due to cation adduction. The vast

majority of these are proton adducts, and the electrospray process produces a distribution of charges, which is thought to be Gaussian in nature [1].

The increase in complexity of data analysis of ESI mass spectra required new algorithms. In 1989 Fenn and coworkers introduced the m/z equation still used today, which incorporates the charge of the ion (z), along with proton mass (m_{proton}) and ion mass (m_{ion}) (Equation 4.1). They simultaneously introduced the first mass spectrum deconvolution algorithm [1].

$$m/z = \frac{m_{ion} + z \cdot m_{proton}}{z} \tag{4.1}$$

In the subsequent years, proteins analysed by ESI-MS have become larger [2] and the number of charges on protein ions have increased concomitantly. In order to calculate masses using Equation 4.1, an iterative computational algorithm is used to determine the correct charge and the mass. As neighbouring peaks have a difference of one charge, a charge state series is assigned for the peaks, e.g. +1,+2,+3. The mass is then calculated for each peak using Equation 4.1 and the standard deviation of all the calculated masses is determined. The masses are then recalculated using one charge higher e.g. +2,+3,+4 and the standard deviation is calculated. This process is repeated and the correct charge state series is assigned to the series which produces the lowest standard deviation of masses. Though this method had been used previously [3], it was formally described by Winkler [4].

E.T. Jaynes introduced the concept of maximum entropy as an optimisation criterion in 1957 [5, 6] and it was applied to mass spectrometry data in 1991 [7]. Maximum entropy was used as a forward algorithm, where a spectrum was deconvoluted assuming that a certain mass was correct [8]. The result of the algorithm would be a stick spectrum of masses, with noise removed, an example of which is shown in Appendix 4.5.1. This algorithm has been kept up to date and is included in the current edition of the MassLynx software package (4.1, Waters Corp.). The algorithm however is not well adapted to large potential mass ranges of analytes as it does not scale well [9] and the "soft" ionisation technique of ESI has progressed from the study of cyclosporin A (1.2 kDa) [10], to native-like analysis of non-covalent complexes over 18

MDa [11].

Larger ions along with higher order oligomeric states led to the acquisition of crowded mass spectra containing overlapping peaks, making analysis difficult. In 2006, a new approach called SOMMS (SOlving complex Macromolecular Mass Spectra) was released [9] and the program introduced the simulation of mass spectra. The parameters to recreate a mass spectrum of multiple components could be entered and a simulated mass spectrum be generated. Manual adjustment of these parameters could create a near exact replica of the original spectrum. This however is extremely time consuming, and though the package has some semi-automatic functionality, it was prone to getting stuck at local minima solutions which resulted in large errors. The program was command line based and, as a consequence of the relatively high knowledge barrier required to operate it, was not widely adopted.

We set out to develop a robust graphical user interfaced program for the analysis of electrospray ionisation mass spectra along with travelling wave ion mobility data. The ESI-MS portion of the program would enable users to manually simulate as with SOMMS but would have a robust automation procedure linked to a graphical user interface (GUI) to ensure easy operation. After we succeeded in developing the simulation/deconvolution algorithm and were developing the ion mobility portion of the program, a different group published the Massign software that had very similar functionality [12, 13]. Massign was developed as a plug-in for the proprietary software LabView (National Instruments, Texas, USA) which now represents an additional financial barrier, and so we have not been able to try it.

Another new mass spectral deconvolution algorithm has been published, named 'Calculating Heterogeneous Assembly and Mass spectra of Proteins' (CHAMP) [14]. This algorithm was designed to deconvolute complex mass spectra where information about the proteins in the spectra are already known, including the mass and accessible surface area (ASA) as taken from X-ray crystallography structures. As the mass of the protein is predetermined, the additional information is used to predict the behaviour of the protein in the mass spectrometer in terms of the level of protonation (estimated using mass) and the number of adduct ions (estimated using mass and ASA).

The presence and proportional abundance of potential complexes are then optimised using non-linear least squares.

4.1.2 Ion mobility mass spectrometry data analysis

Ion mobility is a gas-phase technique that separates ions as they travel under the influence of an applied electric field through a neutral target gas. The time taken for an ion to traverse the cell is related to its mass, charge, and the rotationally averaged collision cross section (CCS) of the ion [15–17]. Ion mobility coupled to mass spectrometry (IM-MS) is a powerful analytical technique that was initially only available in a few laboratories capable of building such specialised instruments. The primary means of performing IM-MS separations was based on drift cell technology [18].

Early drift cell IM-MS experiments were focused towards small proteins and peptides, usually under non-native conditions. Due to the simple ion motion through the drift cell the absolute cross sections of ions could be calculated from the arrival time distributions (ATDs). A typical example of how these data could be represented is shown in Figure 4.1A. For each ion charge state the collision cross section for each peak in the drift time dimension was calculated and plotted. Different charge states can occupy different conformations, with the higher charge states being unfolded, and the lower charge states close to the native conformation.

In 2004 a commercial instrument capable of IM-MS experiments was introduced [20]. The separation of ions was achieved using a travelling wave (T-wave, TW) [21]. In contrast to drift tube instrumentation, a higher level of sensitivity was possible, but the complicated movement of ions during the separation meant that direct calculation of CCS values was not possible [22, 23]. At this time a common way to represent the data was to plot intensity against the arrival time or scans for a single charge state as in Figure 4.1B. To rectify this situation algorithms were developed to calculate analyte CCS values using calibrant proteins of known CCS (see Section 3.2) [24, 25]. One result of the calibration was that now intensity against arrival time plots could be shown as intensity against CCS plots as was often used in drift tube

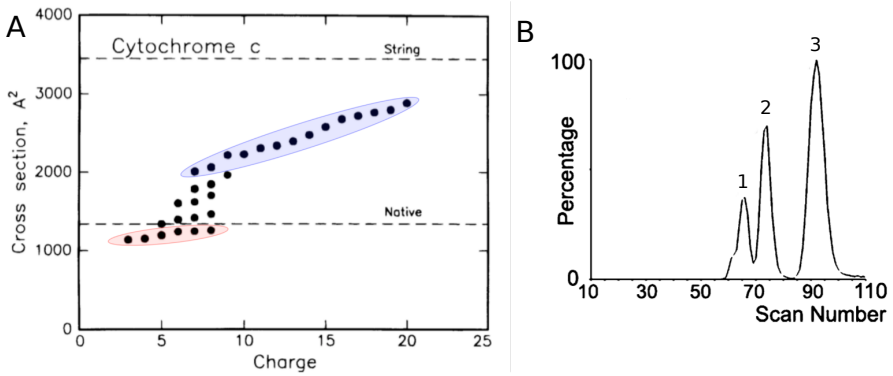

Figure 4.1: (A) Classic ion mobility figure, altered from [19]. Demonstrating the effect of charge state on the collision cross section of cytochrome *c* ions in the gas-phase. Coloured ovals indicate two of the four distinct conformations. Arrival time distribution of bradykinin using TWIM-MS with three distinct protein conformations labelled 1-3 (B). Figure adapted from [20].

experiments [19].

The introduction of commercial ion mobility instruments resulted in an increase in the popularity of the field, however developments in software have been virtually non-existent with the only available software for the interpretation of TWIM-MS data being Driftscope (Waters Corp.). In addition to allowing the plotting of arrival time distributions, Driftscope can produce plots of arrival time against *m/z*. This allows for the full visualisation of the three dimensional data, and an example is shown in Figure 4.2.

The aims of this project were to create a deconvolution algorithm with an easily usable graphical user interface and to link this to ion mobility data analysis. The deconvolution portion of our software presented here, creates a map of which regions of a mass spectrum are associated with which ion, this allows the ATDs to be automatically extracted from the raw data files and is instrumental in the program's ability to easily create IM calibrations that can then be applied to entire data sets automatically. The software can be used to create CCS vs *m/z* heat maps that can be overlaid between different experimental conditions, something that allows for a more in-depth probing of the structural changes taking place between different conditions. Having a program do these analyses allows for the standardisation of the data processing, making the entire process more objective and reproducible between

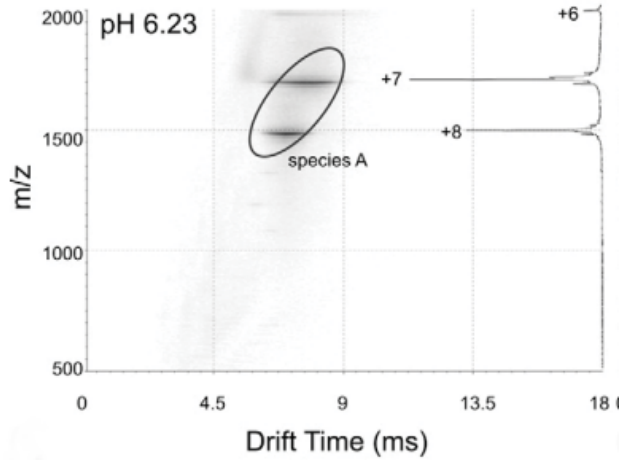

Figure 4.2: An m/z vs. arrival time plot showing the multidimensional data available from a TWIM-MS experiment, with the corresponding mass spectrum displayed as a trace on the right (adapted from [26]).

different practitioners. A number of different uses of the program, with a particular focus, on commonly encountered structural biology applications are illustrated using model proteins.

4.2 Methods

4.2.1 Sample sources

Cytochrome *c* from equine heart, myoglobin from equine heart, alcohol dehydrogenase (ADH) from *Saccharomyces cerevisiae*, bovine serum albumin (BSA), and concanavalin A from *Canavalia ensiformis* were purchased from Sigma Aldrich (St. Louis, MO). Serum amyloid P component (SAP) purified from human serum was purchased from CalBioChem, Merck Biosciences Ltd. (Darmstadt, Germany).

4.2.2 Sample preparation

In order to reduce adduction and the consequent peak broadening, samples must be void of non-volatile salts. In order to maintain native protein struc-

ture the ionic strength of the sample buffer must be maintained. For this reason in the native MS experiments, protein samples were buffer exchanged into 250 mM ammonium acetate. This was achieved by three rounds of dilution-concentration using Amicon Ultra 0.5 ml centrifugal filters (Millipore UK Ltd, Watford UK), with final protein concentration corrected to 20 μM, using a Qubit 2.0 fluorometer.

To further reduce adduction and to obtain more accurate masses, proteins can be denatured. This exposes more amino acid side chains resulting in increased protonation, which in turn increases the electron volts (eV) experienced by the ion in the mass spectrometer. The result of this is reduced adduction and a consequently cleaner mass spectrum. For denaturing experiments, protein samples were buffer exchanged into a 49:49:2 (v:v:v) ratio of H_2O: methanol: acetic acid, and concentrated to 20 μM using Amicon Ultra 0.5 ml centrifugal filters.

4.2.3 Capillary preparation

Nano-ESI compatible capillaries were produced *in house* as described by Hernandez and Robinson [27]. Borosilicate capillary tubes (Harvard Apparatus, Massachusetts, USA) with inner and outer diameter of 0.78 and 1.0 mm respectively were pulled into needles using a Sutter Instrument Co. P-97. To ensure propagation of the electric field the needles were gold coated using an SC7620 sputter coater (Emitech (Quorum), Kent, UK).

4.2.4 nESI-MS calibration

Calibration was performed using 33 mM caesium iodide (Sigma-Aldrich), in 250 mM ammonium acetate. Caesium iodide (CsI) forms a wide array of crystal sizes, whose mass-to-charge ratios (m/z) span beyond the m/z range investigated (500-12,000 m/z). Calibrations were performed by fitting a 5^{th} degree polynomial calibration curve using theoretically determined exact masses of CsI oligomers, using Waters MassLynx software. Spectra obtained were compared and calibrated iteratively until the overall error was less than 10 ppm,

which is checked using the Waters MassDiff software (Figure 4.3).

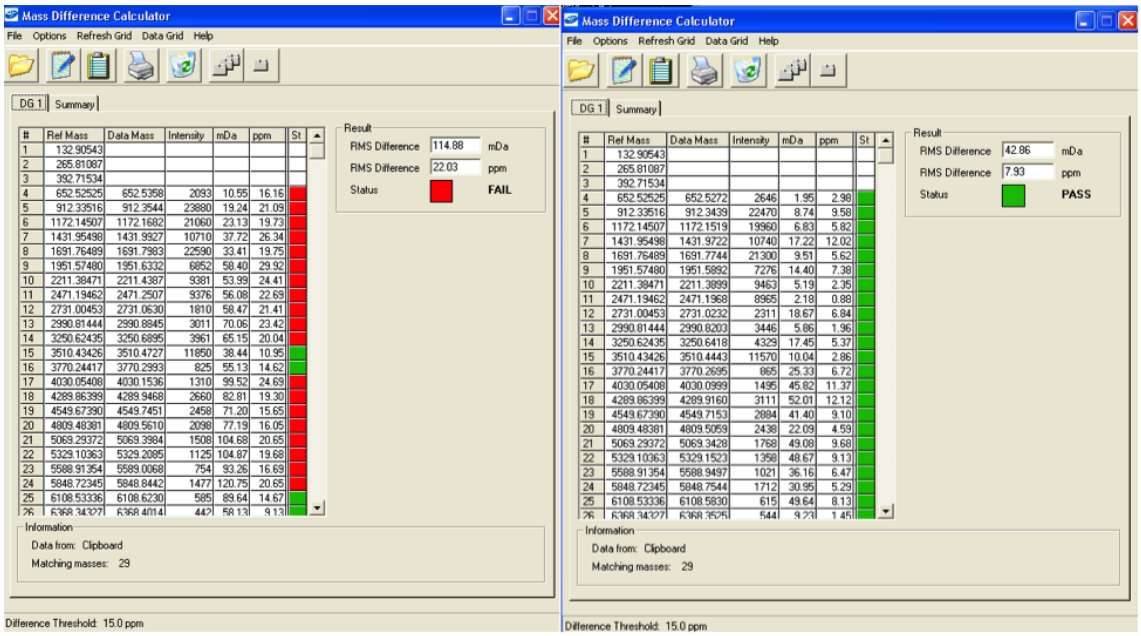

Figure 4.3: Output of the MassDiff program which is used to test if the instrument is properly calibrated. On the left there is an unacceptable result, and the right an acceptable one.

4.2.5 TWIM-MS

Mass spectrometry experiments were carried out on a first generation Synapt HDMS (Waters Corp., Manchester, UK) mass spectrometer [28]. 2.5 μl aliquots of samples were delivered to the mass spectrometer in gold-coated capillaries. Typical instrumental parameters were as follows unless otherwise specified: source pressure 5.5 $m$bar, capillary voltage 1.10 kV, cone voltage 40 V, trap energy 8 V, transfer energy 6 V, bias voltage 15 V, IMS pressure 5.18x10^{-1} $m$bar, IMS wave velocity 250 m/s, IMS wave height 6 V, and trap pressure 4.07x10^{-2} $m$bar. Arrival time to CCS calibration was conducted using the method outlined in Section 3.2.

4.2.6 Experimental procedures

For the heating experiment ADH was incubated at 60 °C for 30 minutes in a heat block. The sample was removed from the heat block and immediately in-

troduced to the mass spectrometer. Instrumental parameters were optimised as follows: source pressure 4.50 $m$bar, cone voltage 60 V, trap energy 15 V, transfer energy 12 V, and IMS wave height 7 V. BSA and concanavalin A were used as CCS calibrants.

For the collision unfolding experiment, the native fold of cytochrome c was disrupted by increasing the bias voltage from 10 V to 80 V at 10 V increments. Instrumental parameters were optimised as follows; source pressure 3.55 $m$bar, cone voltage 30 V, and IMS wave height 7 V. Denatured myoglobin and ADH were used as CCS calibrants.

For the mixing experiment, ADH, BSA, and concanavalin A were mixed in an equimolar ratio, and the instrumental parameters were optimised as follows; trap energy 60 V, transfer energy 30 V, bias voltage 22 V. IMS wave height 7 V. SAP, BSA, and concanavalin A were used as CCS calibrants.

Arsenite oxidase samples were analysed without IM separation using cone and sampling cone voltages of 45 and 2 V, trap and transfer voltages of 15 V and 10 V with a trap pressure of 8.60x10^{-3} $m$bar.

4.2.7 Software development

During a TWIM-MS experiment arrival time distributions (ATDs) are recorded by synchronising the orthogonal acceleration time of flight (oa-ToF) acquisition with the gated release of a packet of ions from the trap T-Wave. For each packet of ions 200 mass spectra are acquired at a rate dependent on the pusher frequency.

The program, Amphitrite, handles the data in the form of an N x 200 matrix (where N is the number of m/z bin increments), with individual vectors to describe the associated axes (i index is associated with the m/z dimension and j with the time dimension). This matrix can be used to generate the full mass spectrum aggregated over arrival time data points ($t_{(i,j)}$) by summing the elements to an N-length vector. Equation 4.2 describes the process of calculating a single m/z value, ξ_i.

$$\xi_i = \sum_{j=0}^{199} t_{(i,j)} \tag{4.2}$$

Additional manipulations can be carried out by selecting sections of the matrix by index, for example each point in the arrival time distribution (t_j) of a particular ion could be extracted by supplying the lower (l) and upper (u) m/z (ξ) index limits, and then summing along the m/z axis as shown in Equation 4.3. The manipulations of this matrix form the basis of the functionality of the program.

$$t_j = \sum_{i=l}^{u} \xi_{(i,j)} \tag{4.3}$$

The software was developed using the Python programming language [29]. Several Python modules were utilised for data analysis including NumPy, SciPy [30] and Matplotlib [31], and the graphical user interface was developed using wxPython [32]. The initial conversion of a raw TWIM-MS file to an Amphitrite project file can only be run under Microsoft Windows as the Waters proprietary library for extracting the data from MassLynx data files is only compatible with Windows. All other aspects of Amphitrite are cross platform and are compatible with GNU/Linux, Mac OS X systems and MS Windows. The software was developed on a workstation with an Intel i7 2600 processor clocked at 3.4 GHz with 16 GB memory, running GNU/Linux - Ubuntu 12.04. Processing times quoted are for a mid-2011 MacBook Air with an Intel i5 1.7 GHz dual-core processor and 4 GB memory. The source code for the project is available at `github.com/gnsiva/amphitrite` and the compiled Windows executables are available from `mscalculator.com` and the Thalassinos lab website*.

*http://www.homepages.ucl.ac.uk/∼ucbtkth/resources.html

4.3 Results and discussion

Until now, Driftscope (Waters Corp.) has been the sole program used to display and manipulate raw TWIM-MS data. The introduction of Amphitrite facilitates increased customisability of plots as well as the automation of previously labour-intensive, subjective and hence non-reproducible tasks. Using model proteins, we describe various examples of how the software can be used.

4.3.1 Mass spectrum simulation

Programs capable of automatic and semi-automatic analysis of mass spectrometry data of proteins and protein complexes are becoming increasingly available [9, 12, 14, 33]. Amphitrite also includes an algorithm for the deconvolution of mass spectra, as fitting peaks to the mass spectrum is the first process in the automatic extraction of corresponding ion ATDs. A Gaussian model (Equation 4.4) [34] is used to represent the distribution of peak heights of the ion peaks within a charge state distribution of a given molecular species, and this is used as a constraint in mass spectral simulations. The individual peaks, corresponding to a specific charge state, can be modelled as Gaussian, Lorentzian (Equation 4.5) [35] or a hybrid peak shape which consists of Gaussian and Lorentzian regions (Equation 4.6) where x is the m/z value, A is the amplitude, μ the mean, and Γ is the full width half maximum (FWHM) of the peak.

$$f(x) = Ae^{-\frac{(\mu-x)^2}{2\left(\frac{\Gamma}{2\sqrt{2\ln 2}}\right)^2}} \tag{4.4}$$

$$f(x) = A\frac{1}{\left[1+\left(\frac{\mu-x}{\Gamma/2}\right)^2\right]} \tag{4.5}$$

$$f(x) = \begin{cases} Ae^{-\frac{(\mu-x)^2}{2\left(\frac{\Gamma}{2\sqrt{2\ln 2}}\right)^2}} & : x \leq \mu \\ A\frac{1}{\left[1+\left(\frac{\mu-x}{\Gamma/2}\right)^2\right]} & : x > \mu \end{cases} \tag{4.6}$$

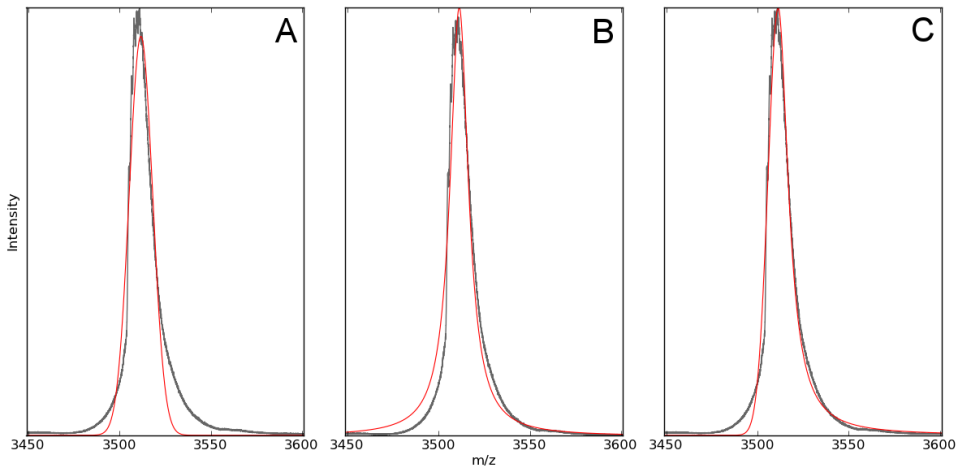

Figure 4.4: A typical nESI protein charge state peak (grey). Overlaid are three peak models whose parameters were optimised to match the experimental data; A: Gaussian, B: Lorentzian and C: hybrid.

A demonstration of the three peak models is shown in Figure 4.4, in each case the fit for the experimental data was optimised using non-linear least squares. Throughout this chapter the hybrid peak model has been used and as a result the model for a single charge state series is described by Equation 4.7. z_0 and z_n represent the lowest and highest charge state in the series, A_z, μ_z and Γ_z represent the amplitude, mean and FWHM parameters for the charge state series Gaussian distribution respectively and H^+ is the mass of a proton. Additionally the mass has been denoted as "mass" to distinguish $\frac{\text{mass}}{z}$ from mass-to-charge ratio (m/z), as readout by the mass spectrometer.

$$f(x) = \sum_{z_i=z_0}^{z_n} A_z e^{-\frac{\left(\left(\frac{\text{mass}}{z_i}+\text{H}^+\right)-\mu_z\right)^2}{2\cdot\left(\frac{\Gamma_z}{2\sqrt{2\ln 2}}\right)^2}} \cdot \begin{cases} e^{-\frac{(\mu-x)^2}{2\cdot\left(\frac{\Gamma}{2\sqrt{2\ln 2}}\right)^2}} & : x \leq \mu \\ \frac{1}{\left[1+\left(\frac{\mu-x}{\Gamma/2}\right)^2\right]} & : x > \mu \end{cases} \tag{4.7}$$

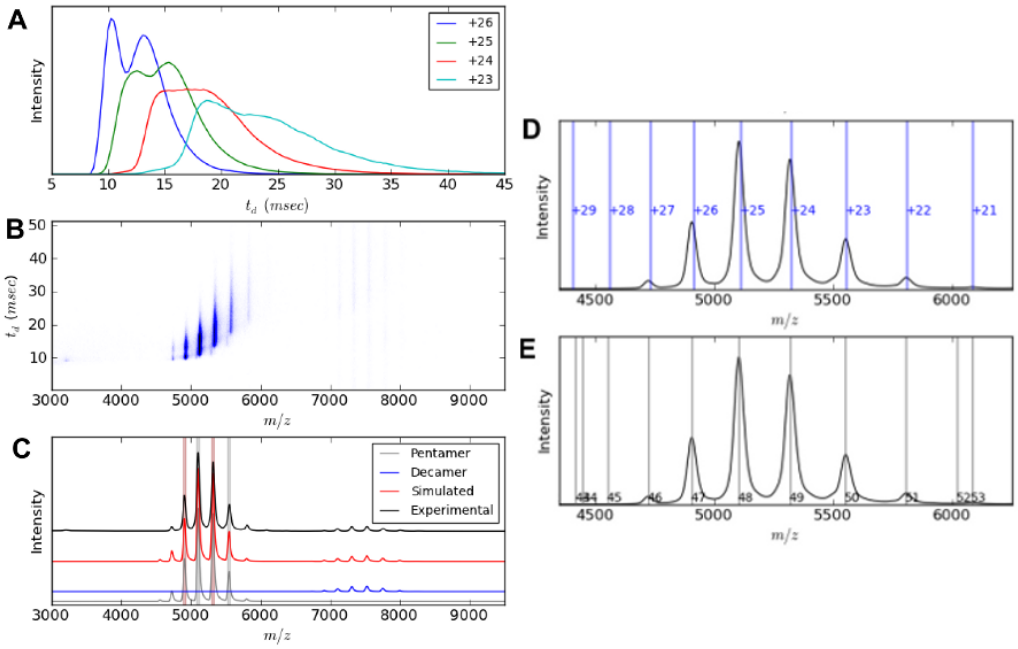

Figure 4.5: Different stages in extracting arrival time distribution plots of serum amyloid P (pentamer). The user selects peaks of the charge state series in (E) (numbers are unique peak identifiers), the mass is calculated and the theoretical charge states are then plotted over the spectrum (panel D), with the subsequent simulated spectrum plotted in (C). (B) Mass mobility plot, using the simulated spectrum each charge state can be identified and the ATDs extracted. The ATDs have been displayed as overlaid ATD distributions for each charge state as shown in (A).

After minimal user input the program can simulate a mass spectrum as presented in Figure 4.5C with a computational processing time of under 2 seconds. To assess the quality of the fit an error statistic (E) is calculated as the mean deviation using Equation 4.8, where f is the function in Equation 4.7, y is the experimental data, N is the total number of data points, s is the space between m/z data points and A_z,Γ_z,μ_z and Γ are the parameters output by the fitting procedure. In the case of Figure 4.5C, the error was 0.47 % (of base peak intensity) per m/z.

$$E = \sum_{i=0}^{N} \frac{|f(x_i, A_z, \Gamma_z, \mu_z, \Gamma) - y|}{N \cdot s \cdot \max y} \tag{4.8}$$

There are two ways in which one can specify the input required. If the expected mass of the macromolecule is known, it can be manually entered,

along with the charge state range over which to simulate that particular mass e.g. +22 to +27 (Figure 4.5D). More than one mass can be entered and after this, the program uses the non-linear least squares optimisation algorithm [36] to minimise the difference between the simulated and experimental data. If, however, the mass of the components in the spectrum is unknown, the program aids the user in this process. The program uses the gradient to automatically identify peaks (where $f'(m/z) = 0$ and $f''(m/z) < 0$) in the mass spectrum which are then given arbitrary unique numerical identifiers as shown in Figure 4.5E.

The user then selects the sequential charge state peaks of a particular species using the numerical peak identifiers. The mass of the species is calculated using the m/z values of the peak tops. The theoretical m/z values for charge states are calculated (default +1 to +100) and are displayed as vertical markers along with the calculated mass and error (Figure 4.5D). Both of these features help to ensure that peaks were correctly identified, as incorrect peak picking would result in misaligned theoretical charge states and large mass errors. The program automatically estimates the charges to simulate, which the user can edit. After this process has been completed for each species, the program can fit simulated data to the supplied spectrum using least squares optimisation with the result shown in Figure 4.5C. If the user then notices that a species was missed, it can be added to the simulation by repeating the steps described above. The data simulation algorithm can identify and deconvolve overlapping peaks and peak shoulders (see example in Figure 4.8C).

4.3.2 ATD extraction

Standard m/z against arrival time (t_d) plots like those displayed by Driftscope can be drawn by Amphitrite (as previously shown in Figure 4.5B) and a key improvement is the resolution of these images. In Amphitrite, the user can determine the space between each data point in the m/z space i.e. how wide a particular m/z bin is. For the figures shown here a spacing of 2 m/z units was used.

Extracting ATDs across all charge states of a given spectrum has now

been streamlined as the fitting procedure previously described determines the FWHM and peak centre of each of the peaks in the mass spectrum, and uses this information to automatically extract the corresponding ATD for each charge state, with the results shown in Figure 4.5C.

In experiments where multiple spectra are obtained of the same protein under different conditions, the ATDs corresponding to a single charge state can be extracted for all the files in a similar manner as exemplified in Figure 4.11.

4.3.3 Calibration

Protocols to convert TWIM-MS arrival times to CCS have been described previously [24, 25, 37, 38], and Amphitrite automates the calibration procedure outlined in Section 3.2, thereby reducing subjectivity that can be introduced during the ATD extraction and subsequent t_d peak selection.

The process for creating a calibration is shown in Figure 4.6. From a user input perspective, the program is given the calibrant raw data file and the name of that calibrant, in this case myoglobin. Creating a calibration with more than one calibrant protein is also possible. It then automatically selects the charge states (vertical bands in Figure 4.6A) that have corresponding published CCSs [39]. Low abundance charge state peaks can be deselected and ignored in order to improve the fit. The program automatically takes t_d value corresponding to the highest intensity to use in the calibration and produces the output shown in Figure 4.6C. Another way in which outliers can be addressed, is that the program allows the user to move the arrival time used in the calibration from the highest intensity point, this can help with skewed distributions and noise.

Outliers can be addressed by specifying alternate peak apexes (which are also automatically detected), by specifically providing a t_d value as input, or by removing the peak from the calibration.

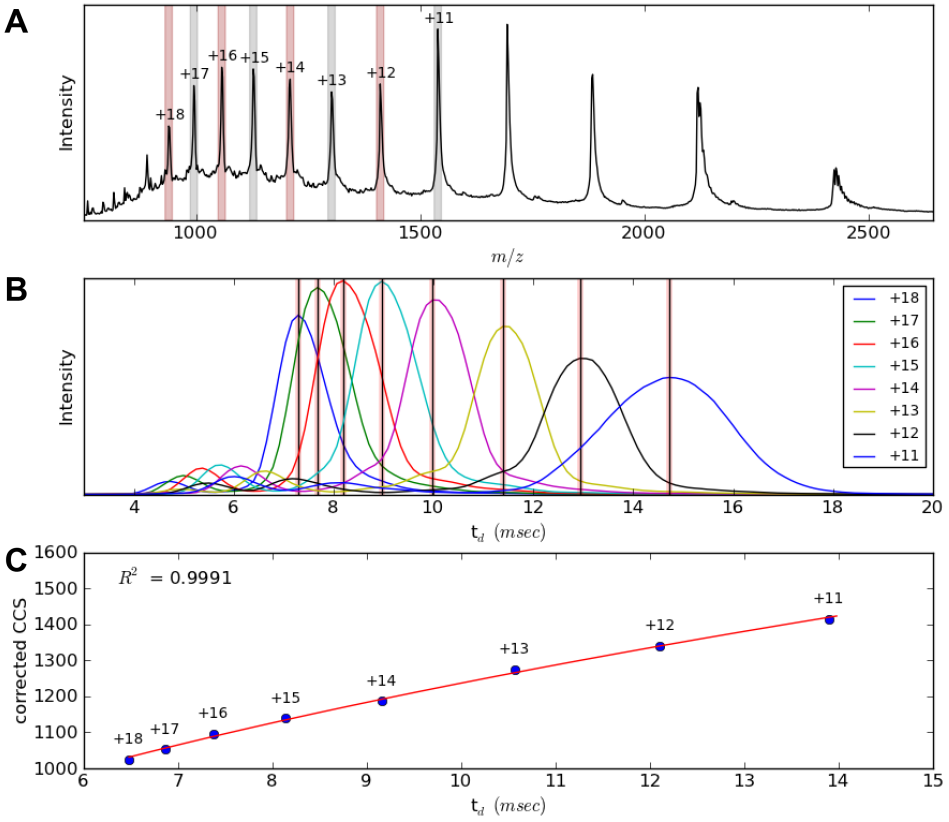

Figure 4.6: Creation of a CCS calibration using denatured myoglobin. Amphitrite automatically selects charge states (A), which correspond to published CCS†. From the selected peaks, the ATDs are extracted and plotted (B). The peak tops are automatically picked and displayed. A calibration curve, using a power fit to the data, is then calculated and plotted. Poor fits can be recalculated by manually adjusting the peak tops selected in the previous stage. The calibration procedure used has been described in [24].

4.3.4 Applying a calibration

The ability to read the raw data files has allowed a more fine-grained approach to applying a calibration to TWIM-MS data. To illustrate this method, the calibration equations from Section 3.2 have been simplified to create Equation 4.9 (see Appendix 4.5.2 for derivation), where t is the drift time and ξ is the m/z value.

$$\Omega = \frac{A \cdot z}{\sqrt{\mu}} \cdot \left(t - \frac{276}{V_w} - 0.085 \cdot \sqrt{\frac{\xi}{1000}} \right)^B \tag{4.9}$$

The existing method for calibrating TWIM-MS data is described in Equation 4.10. Here the drift time is represented as a summed portion of the data matrix as described in Equation 4.3, m/z value's index is k corresponding to the m/z value of the centre of the peak in the mass spectrum.

$$\Omega = \frac{A \cdot z}{\sqrt{\mu}} \cdot \left(\left(\sum_{i=l}^{u} \xi_{(i,j)} \right) - \frac{276}{V_w} - 0.085 \cdot \sqrt{\frac{\xi_k}{1000}} \right)^B \tag{4.10}$$

The new calibration method (Equation 4.11) however, allows each t_d value to be associated throughout the calculation with the correct m/z value, thereby reducing error caused by only using a single m/z value to indicate the m/z range covered by a mass spectrum peak.

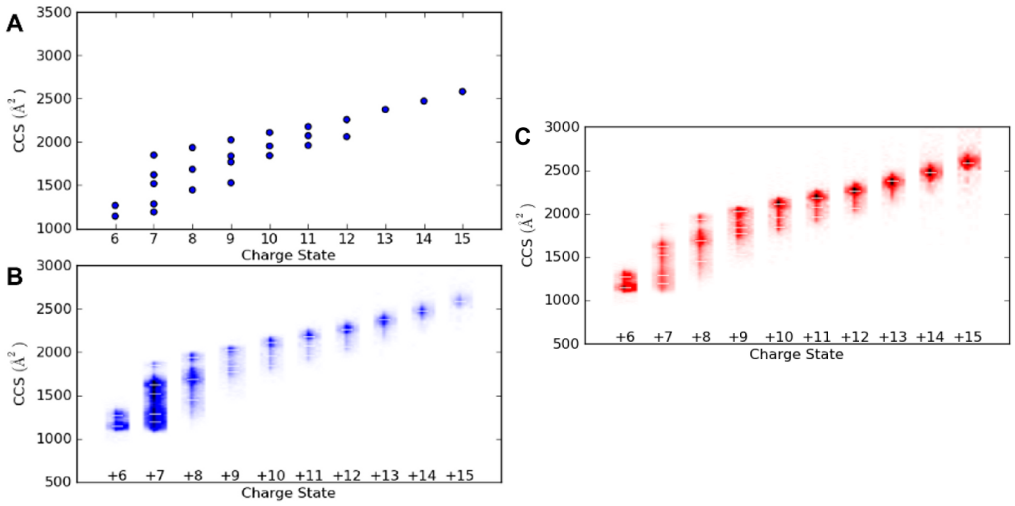

Figure 4.7: Charge state collision cross section plots of cytochrome c. A calibration similar to that shown in Figure 4.6 was applied to the t_d data to generate the CCS data. A dot in panel A represents each peak top in the t_d for that particular charge state. The same data are shown in panel B as a heat map with peak intensity being represented by the colour intensity and the CCS of peak tops shown as a white dash. Panel C shows these data normalised by individual peak volume.

$$\Omega = \sum_{i=l}^{u} \frac{A \cdot z}{(u-l) \cdot \sqrt{\mu}} \left(\xi_{(i,j)} - \frac{276}{V_w} - 0.085 \cdot \sqrt{\frac{\xi_i}{1000}} \right)^B \tag{4.11}$$

In Figure 4.7A an archetypal IM-MS experiment result is shown, CCS vs. charge state for denatured cytochrome c. This is a common way found in the

literature to present IM-MS data, however, a lot of the original information in the data is lost. Where more than one conformation exists for a given charge state, such as for the +8 charge state shown in Figure 4.7, there is no way of deducing the relative intensities of the different conformations. It is also not possible to infer the width of each conformation in the CCS dimension. Biologically, this can be very informative as an increase in the width of a CCS distribution can indicate increased conformational flexibility [12, 40] and as shown recently can, in certain cases, also limit the observed ATD resolution of higher resolving instruments [41]. Comparing two proteins by overlaying figures like those shown in Figure 4.7A can miss important conformational changes as the peak CCS can remain the same, while the width of the CCS distribution can change between two conditions [40]. The figures generated by Amphitrite (Figures 4.7B and C) also provide visual information regarding the width of the MS peak (in the x-axis direction) that was used to reconstruct the ATDs. The program displays the CCS dimension peak tops, as shown in Figure 4.7A, and these are calculated automatically (where $f'(\Omega) = 0$ and $f''(\Omega) < 0$).

Features of low abundance can be visualised by using a different method of normalising the peak intensities. In Figure 4.7B the colour saturation is normalised to the intensity of the base peak in the mass spectrum, i.e. to the maximum intensity in the entire matrix holding the data, so that one can see that the +7 charge state is the most intense peak in the mass spectrum. In Figure 4.7C the intensity of each charge state is normalised to the total intensity for that m/z slice. This allows for less abundant features to be visualised by increasing the dynamic range of the display for lower abundance charge states and for the conformational flexibility and adduction to be readily assessed.

4.3.5 Complex mixture analysis

The program's mass spectral deconvolution algorithm can be coupled with the modified calibration approach described in Section 3.2 to analyse complex ion mobility spectra, and to make sense of the complicated t_d vs m/z plot

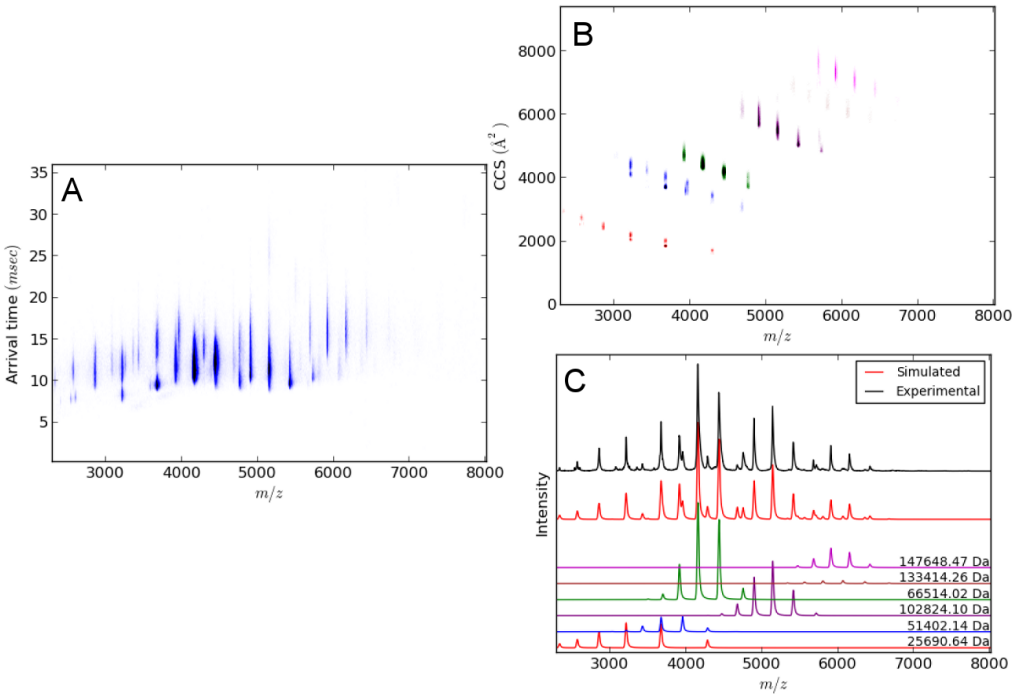

Figure 4.8: IM-MS analysis of a mixture of BSA, concanavalin A and alcohol dehydrogenase. The mass spectrum was deconvoluted into its component parts (C), with the raw arrival time distribution shown in panel A. Using the deconvolution data and CCS calibration (like that shown in Figure 4.6), the raw arrival times can be separated and converted into CCS vs. m/z information for each molecular component (B). The colouring is consistent between panels A and B (concanavalin A monomer - red, dimer - blue, tetramer - purple, BSA monomer - green, dimer - brown, ADH tetramer - magenta).

produced by Driftscope.

As seen in Figure 4.8C the program successfully deconvolutes the mass spectrum by identifying and calculating the mass and the charge state distribution parameters for all species. Additionally the individual ion peak widths are determined (as in Figure 4.8C). The typical t_d vs. m/z plot is shown in Figure 4.8A and when observed in isolation does not easily allow one to identify the number of species present. Using the parameters determined in Figure 4.8C, a calibration like the one shown in Figure 4.6 can be applied, transforming the t_d vs. m/z plot into a CCS vs. m/z plot (Figure 4.8B). This conversion into absolute cross section separates out the individual species and results in a plot which can be more readily analysed.

4.3.6 Spectral averaging

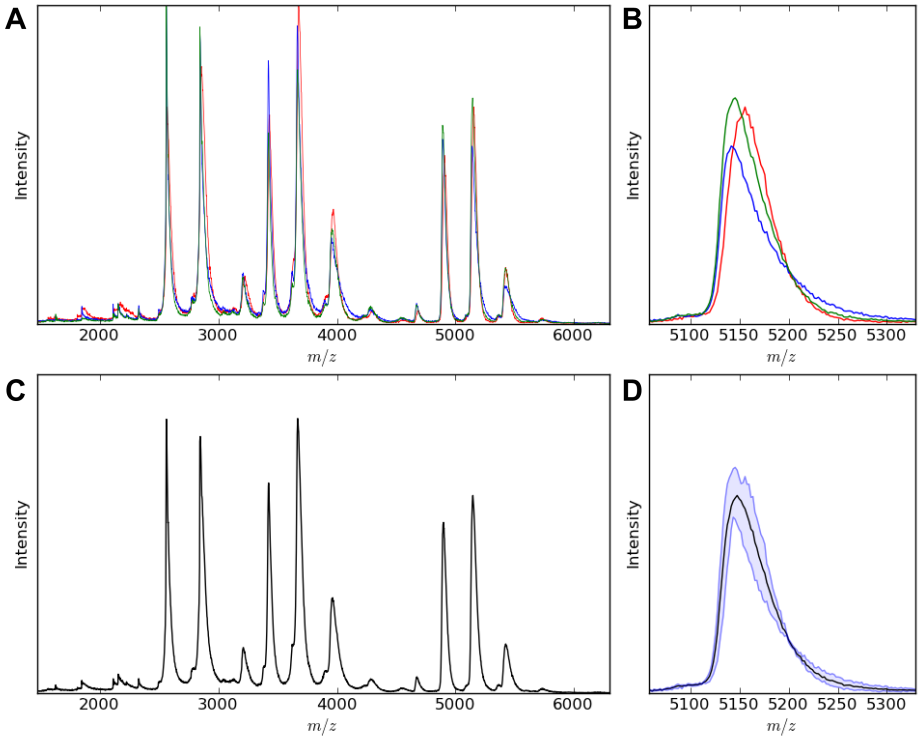

Figure 4.9: Spectral averaging. Panel A, shows the variation for the same sample measured using different capillary needles. Each individual spectrum has been overlaid and coloured differently. A peak at approximately 5,150 m/z has been enlarged to portray more clearly the variation between each experiment (panel B). The average of the three spectra in panel A is plotted in panel C. The same enlarged peak in panel B is again shown in panel D, with the minimum and the maximum of the three spectra plotted as light blue lines and the mean plotted in black.

Collecting multiple mass spectra can help reduce the error and electrospray variation caused by certain factors such as needle-to-needle variation and needle positioning. Figure 4.9A and B show the effect of obtaining three mass spectra of the same sample using the same instrumental conditions but with three different needles. From the analysis of mass spectra of samples under different conditions (e.g. temperature), peak intensities and areas can offer information additional to the mass of the ions.

By comparing the integrals of different oligomeric species under differing conditions (e.g. temperature), the formation or disappearance of oligomers in response to conditions of interest can be inferred. It is advantageous to be

able to average technical replicates and compare those between conditions in order to assess whether changes are due to differing conditions rather than technical variability.

Using the program one can average spectra in the mass spectrum, t_d and CCS space. Figure 4.9D shows the band between minimum and maximum intensities for a peak, with the average in the centre. The program can also be set to display the error range in terms of standard deviation, quartile ranges and percentage errors. Peak areas and heights can be automatically extracted to be used in further analyses.

4.3.7 Comparing different conditions

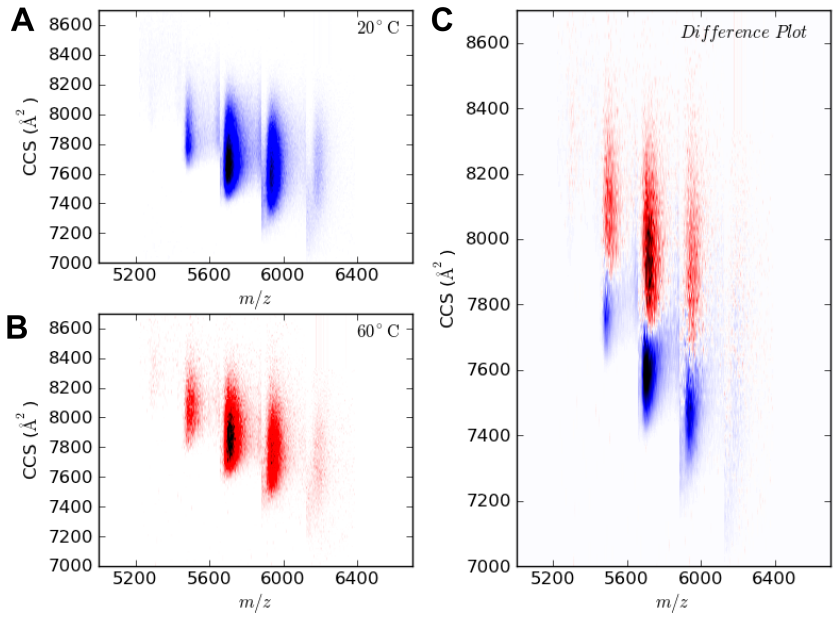

Figure 4.10: Comparison of heating experiments of alcohol dehydrogenase at 20 °C and 60 °C. Data at each temperature was replicated and averaged as shown in figure 4.9. Panel A and B show the distribution of collision cross sections for 20 °C and 60 °C respectively. A difference plot is shown in panel C, where overlapping CCS distributions in pane A and B have been subtracted from one another.

Spectral averaging was performed on the heating experiment spectra used in Figure 4.10. IM-MS can be used to monitor the effect on conformation of a discreet stressor such as heating [42]. To demonstrate this, ADH spectra were acquired with the protein at room temperature (20 °C), and after heating

at 60 °C for 30 minutes. For each condition three technical replicates were acquired. In addition to the CCS vs. *m/z* plot of the sample at 20 °C (Figure 4.10A) and 60 °C (Figure 4.10B), a difference plot can also be drawn. This is shown in Figure 4.10C, with the colours matching those used in Figure 4.10A and B. In this figure the 20 °C and 60 °C spectra have been normalised to the volume within the plot, as this helps to make the comparison more representative. The data show that the process of heating ADH causes it to adopt a more open conformation as demonstrated by the increase in CCS.

4.3.8 Collision induced unfolding

A unique and very informative IM-MS experiment is gas-phase, collision-induced unfolding which can be achieved by accelerating ions into the gas filled region, prior to their entry into the T-Wave mobility cell (Figure 4.11). This experiment can probe the unfolding of protein ions by monitoring changes in ATDs/CCSs upon increasing collision energies [43–46] and is covered in more detail in Chapter 5. This experiment can be used to probe a number of experimental conditions and their effect on protein conformation/stability, however, it can generate a large number of data files as each time the collision energy is changed a different raw data file is recorded. The automation offered by Amphitrite greatly facilitates the processing of such datasets.

Figure 4.11 shows the results of the collision induced unfolding for the +6 charge state of cytochrome *c*. Data were recorded at eight different collision energies. The program uses the data file which was acquired at the lowest collision energy to identify and perform a fit (in order to calculate a FWHM for that peak) to the mass spectrum as previously described in the materials and methods section. Alternatively, explicit mass ranges can also be entered manually. The *m/z* range is then used to automatically extract ATDs from all files in the dataset. If a CCS calibration is provided at this stage, all ATDs are converted to CCS values. The reduced amplitude for the peaks seen for 10 and 20 V are due to the peak broadening effects of bound adducts which lessens as the collision energy is increased. This highlights the benefit of the new plot (Figure 4.11A); changes in both the *m/z* and the ATD / CCS dimensions can

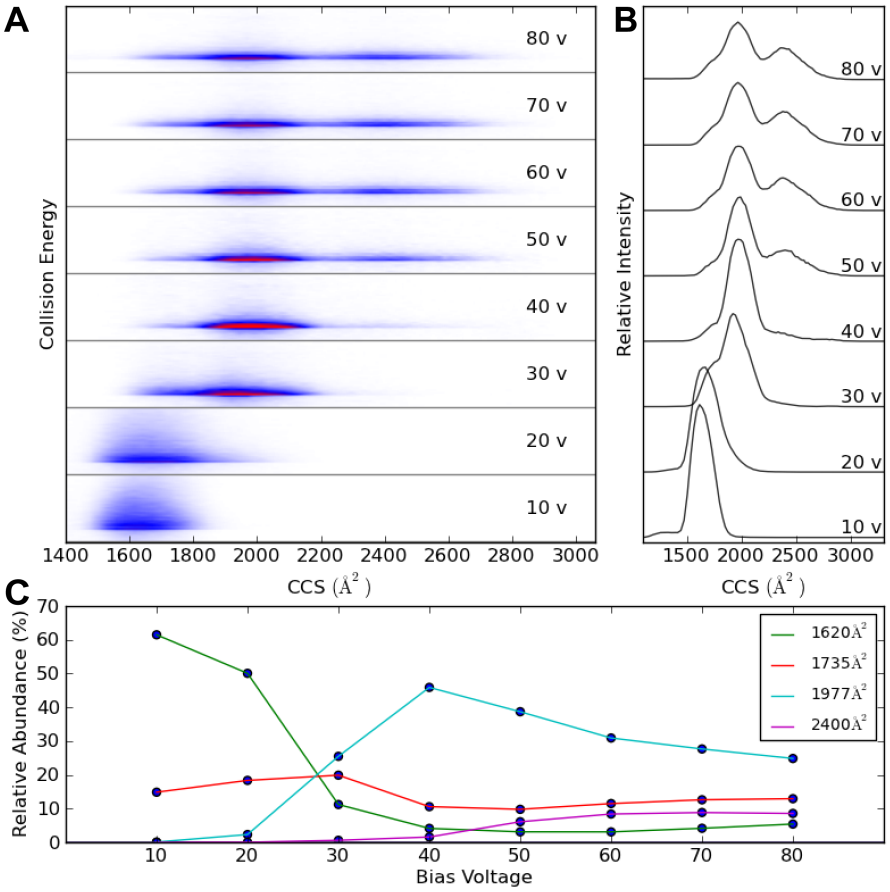

Figure 4.11: Collision induced unfolding of cytochrome *c*. The CCS distribution of the 6+ charge state is shown in panel A as the collision energy is increased from 10 V to 80 V at 10 V increments. Intensities for panel A have been normalised to the total ion intensity for each three dimensional peak. The corresponding CCS plot for each voltage increment is shown in panel B. The relative intensity of each peak top identified from the CCS plots is also monitored as the bias voltage is increased (panel C).

be visualised simultaneously providing a means of globally monitoring ion structural changes during collision induced unfolding experiments. Finally, the program can track the peak intensities of given CCS or t_d values as shown in Figure 4.11C.

4.3.9 Arsenite oxidase

To further illustrate the benefits of the new software some data from research into arsenite oxidase will be shown. In order to investigate the constituents of

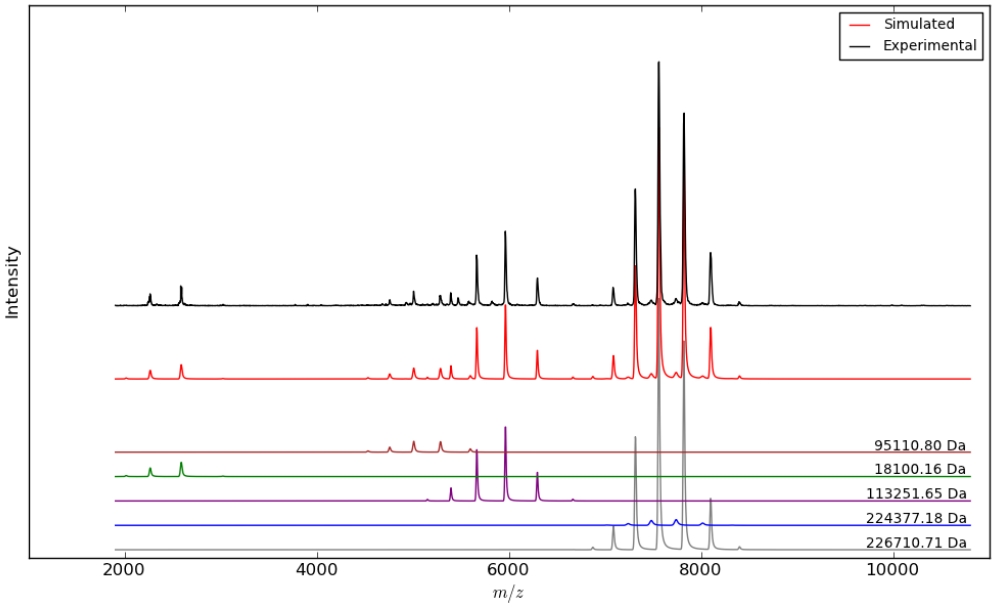

Figure 4.12: Deconvolution of an arsenite oxidase mass spectrum.

the purified protein the spectrum in Figure 4.12 was collected. The protein is a tetramer made up of two heterodimers (AioA-AioB). The sample is further complicated as each subunit in the heterodimer has a different iron-sulphur cluster and each the dimer is also associated with a cofactor. Amphitrite successfully deconvoluted the initial complex spectrum, assigning masses to its constituents and confirming the purity of the sample.

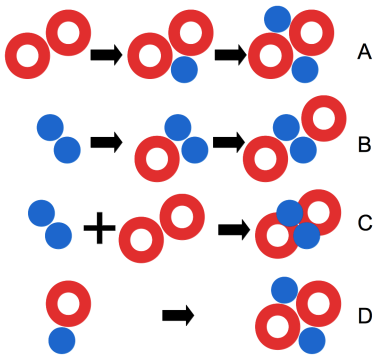

Figure 4.13: Diagram of the four potential assembly pathways of arsenite oxidase. Blue circles represent subunit AioB and red rings represent AioA.

An X-ray crystallography model was created for the protein, however the model does not define the assembly pathway of the complex. The potential pathways that the protein could form are shown in Figure 4.13 [47]. Mass

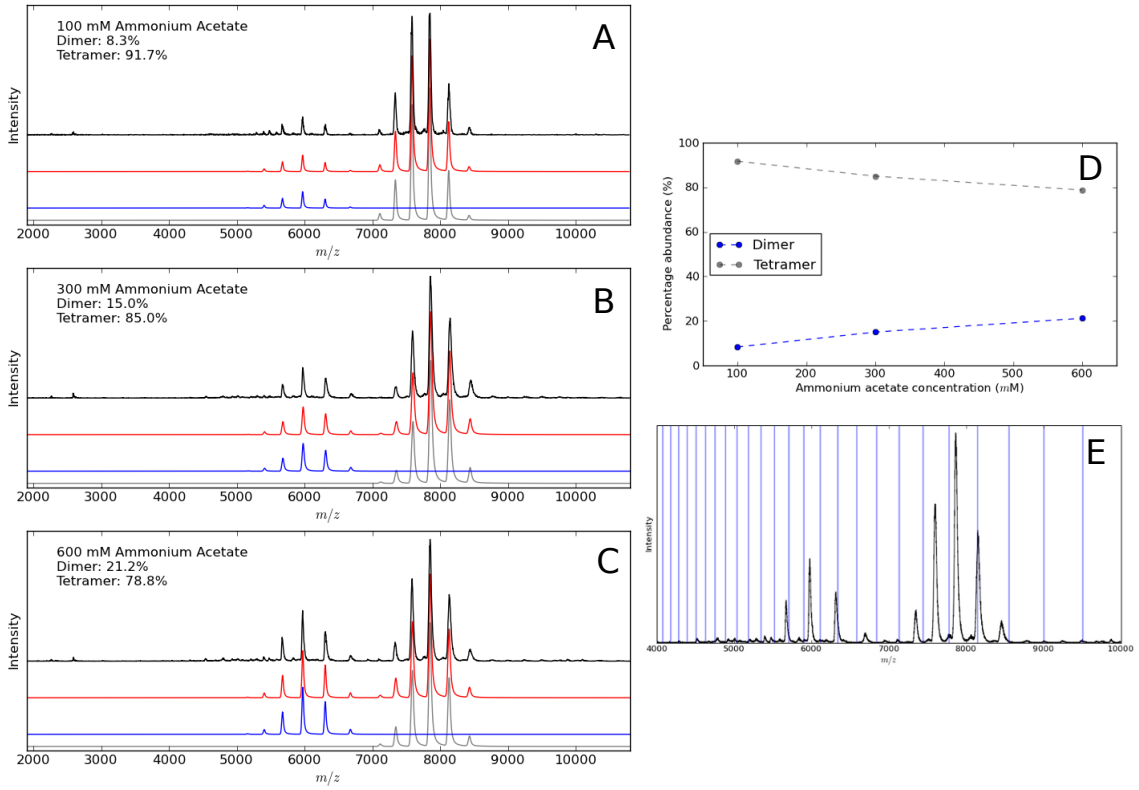

Figure 4.14: (A,B,C) Assembly pathway analysis of arsenite oxidase. Investigating the change in relative abundance of oligomeric states of arsenite oxidase in relation to ionic disruption by means of ammonium acetate concentration. Dimer component trace is blue, and tetramer is grey. (D) Summary of the integrals. (E) The program indicating the theoretical charge states of the mass of a potential trimer.

spectrometry can be used to probe the assembly pathway by altering the ionic strength of the buffer and observing the resultant oligomer distribution. The ammonium acetate concentration was increased (Figure 4.14A, B and C) and a simplified deconvolution was performed using Amphitrite to track the proportional abundance of the two most common species by calculating their integrals. As the ammonium acetate concentration increased the proportion of heterodimer (blue trace) increased, this demonstrates that increased ionic stress causes a larger proportion of the protein to exist as a heterodimer. Pathways A and B are disproven by the absence of any peaks at trimer m/z values (Figure 4.14E), similarly pathway C is not as no peaks were found to indicate the mass of either homodimer.

The software allowed the checking of the purity and validity of the protein

expression and purification. It was also used to track the change in abundance of differing species in response to external stimuli and was able to check for the presence of species of given mass by displaying the theoretical m/z values.

4.3.10 Current state of software development

The code for the Amphitrite project is open source and available from `github.com/gnsiva/Amphitrite`. The code is well documented, and users can create their own scripts to analyse their data.

Software package also contains prototype graphical user interfaces (GUIs) which are able to carry out all of the analyses shown in this Chapter, a screenshot of each of these is provided in Appendix 4.5.3. All of the GUIs with the exception of SpectralAveragingGui have been compiled to Windows binaries that are available on `mscalculator.com`.

It is hoped that this project will be continued by hiring full-time software developers to further develop the GUIs. In addition my supervisor has indicated that he would take the responsibility of creating tutorials (webpage or video) to help users with using Amphitrite.

4.4 Conclusion

Amphitrite introduces a simple, robust and reproducible method for deconvoluting complex spectra. The simulations performed have the added advantage of being able to estimate peak area and so the relative abundance of molecular components as well as determining their mass and charge state distribution parameters.

The fitting functionality introduced substantially enhances the processing of TWIM-MS data, making the process automated and less prone to user subjectivity. It also allows for a more detailed analysis of the data acquired and the automatic comparison of entire TWIM-MS datasets between different experimental conditions. The data presented here demonstrated common applications in structural mass spectrometry as well as new ways of representing and analysing the data. The developments facilitate the analyses required to interpret the native MS experimental data throughout the Ph.D.

Contributions

All experiments were designed and analysed by Ganesh N. Sivalingam. TWIM-MS data for Figures 4.5, 4.7, 4.10, and 4.11 was collected by Dr. Jun Yan. Harpal Sahota acquired the data for Figure 4.8, with the remaining data being acquired by Ganesh N. Sivalingam.

References

[1] Mann, M., Meng, C. K., and Fenn, J. B. (1989). "Interpreting mass spectra of multiply charged ions". *Analytical Chemistry* 61.15, pp. 1702–1708.

[2] Uetrecht, C., Barbu, I. M., Shoemaker, G. K., Duijn, E. van, and Heck, A. J. (2011). "Interrogating viral capsid assembly with ion mobility–mass spectrometry". *Nature Chemistry* 3.2, pp. 126–132.

[3] Tito, M. A., Tars, K., Valegard, K., Hajdu, J., and Robinson, C. V. (2000). "Electrospray time-of-flight mass spectrometry of the intact MS2 virus capsid". *Journal of the American Chemical Society* 122.14, pp. 3550–3551.

[4] Winkler, R. (2010). "ESIprot: a universal tool for charge state determination and molecular weight calculation of proteins from electrospray ionization mass spectrometry data". *Rapid Communications in Mass Spectrometry* 24.3, pp. 285–294.

[5] Jaynes, E. T. (1957a). "Information theory and statistical mechanics". *Physical Review* 106.4, p. 620.

[6] Jaynes, E. T. (1957b). "Information theory and statistical mechanics II". *Physical Review* 108.2, p. 171.

[7] Ferrige, A. G., Seddon, M. J., Jarvis, S., Skilling, J., and Aplin, R. (1991). "Maximum entropy deconvolution in electrospray mass spectrometry". *Rapid Communications in Mass Spectrometry* 5.8, pp. 374–377.

[8] Ferrige, A. G., Seddon, M. J., Green, B. N., Jarvis, S. A., Skilling, J., and Staunton, J. (1992). "Disentangling electrospray spectra with maximum entropy". *Rapid Communications in Mass Spectrometry* 6.11, pp. 707–711.

[9] Breukelen, B. van, Barendregt, A., Heck, A. J., and Heuvel, R. H. van den (2006). "Resolving stoichiometries and oligomeric states of glutamate synthase protein complexes with curve fitting and simulation of

electrospray mass spectra". *Rapid Communications in Mass Spectrometry* 20.16, pp. 2490–2496.

[10] Whitehouse, C. M., Dreyer, R. N., Yamashita, M., and Fenn, J. B. (1985). "Electrospray interface for liquid chromatographs and mass spectrometers". *Analytical Chemistry* 57.3, pp. 675–679.

[11] Snijder, J., Rose, R. J., Veesler, D., Johnson, J. E., and Heck, A. J. R. (2013). "Studying 18 MDa Virus Assemblies with Native Mass Spectrometry". *Angewandte Chemie International Edition* 52.14, pp. 4020–4023.

[12] Zhou, M., Morgner, N., Barrera, N. P., Politis, A., Isaacson, S. C., Matak-Vinković, D., Murata, T., Bernal, R. A., Stock, D., and Robinson, C. V. (2011). "Mass spectrometry of intact V-type ATPases reveals bound lipids and the effects of nucleotide binding". *Science* 334.6054, pp. 380–385.

[13] Morgner, N. and Robinson, C. V. (2012). "Massign: An assignment strategy for maximizing information from the mass spectra of heterogeneous protein assemblies". *Analytical Chemistry* 84.6, pp. 2939–2948.

[14] Stengel, F., Baldwin, A. J., Bush, M. F., Hilton, G. R., Lioe, H., Basha, E., Jaya, N., Vierling, E., and Benesch, J. L. (2012). "Dissecting heterogeneous molecular chaperone complexes using a mass spectrum deconvolution approach". *Chemistry & Biology* 19.5, pp. 599–607.

[15] Jarrold, M. F. (2000). "Peptides and proteins in the vapor phase". *Annual Review of Physical Chemistry* 51.1, pp. 179–207.

[16] Hoaglund-Hyzer, C. S., Counterman, A. E., and Clemmer, D. E. (1999). "Anhydrous protein ions". *Chemical Reviews* 99.10, pp. 3037–3080.

[17] Helden, G. von, Wyttenbach, T., and Bowers, M. T. (1995). "Conformation of macromolecules in the gas phase: use of matrix-assisted laser desorption methods in ion chromatography". *Science* 267.5203, pp. 1483–1485.

[18] Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., and Hill, H. H. (2008). "Ion mobility–mass spectrometry". *Journal of Mass Spectrometry* 43.1, pp. 1–22.

[19] Clemmer, D. E. and Jarrold, M. F. (1997). "Ion mobility measurements and their applications to clusters and biomolecules". *Journal of Mass Spectrometry* 32.6, pp. 577–592.

[20] Thalassinos, K., Slade, S. E., Jennings, K. R., Scrivens, J. H., Giles, K., Wildgoose, J., Hoyes, J., Bateman, R. H., and Bowers, M. T. (2004). "Ion mobility mass spectrometry of proteins in a modified commercial mass spectrometer". *International Journal of Mass Spectrometry* 236.1, pp. 55–63.

[21] Giles, K., Pringle, S. D., Worthington, K. R., Little, D., Wildgoose, J. L., and Bateman, R. H. (2004). "Applications of a travelling wave-based radio-frequency-only stacked ring ion guide". *Rapid Communications in Mass Spectrometry* 18.20, pp. 2401–2414.

[22] Henderson, S. C., Valentine, S. J., Counterman, A. E., and Clemmer, D. E. (1999). "ESI/ion trap/ion mobility/time-of-flight mass spectrometry for rapid and sensitive analysis of biomolecular mixtures". *Analytical Chemistry* 71.2, pp. 291–301.

[23] Shvartsburg, A. A. and Smith, R. D. (2008). "Fundamentals of traveling wave ion mobility spectrometry". *Analytical Chemistry* 80.24, pp. 9689–9699.

[24] Thalassinos, K., Grabenauer, M., Slade, S. E., Hilton, G. R., Bowers, M. T., and Scrivens, J. H. (2009). "Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry". *Analytical Chemistry* 81.1, pp. 248–254.

[25] Ruotolo, B. T., Benesch, J. L., Sandercock, A. M., Hyung, S.-J., and Robinson, C. V. (2008). "Ion mobility–mass spectrometry analysis of large protein complexes". *Nature Protocols* 3.7, pp. 1139–1152.

[26] Smith, D. P., Giles, K., Bateman, R. H., Radford, S. E., and Ashcroft, A. E. (2007). "Monitoring copopulated conformational states during protein folding events using electrospray ionization-ion mobility spectrometry-mass spectrometry". *Journal of the American Society for Mass Spectrometry* 18.12, pp. 2180–2190.

[27] Hernández, H. and Robinson, C. V. (2007). "Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry". *Nature Protocols* 2.3, pp. 715–726.

[28] Pringle, S. D., Giles, K., Wildgoose, J. L., Williams, J. P., Slade, S. E., Thalassinos, K., Bateman, R. H., Bowers, M. T., and Scrivens, J. H. (2007). "An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument". *International Journal of Mass Spectrometry* 261.1, pp. 1–12.

[29] Van Rossum, G. and Drake, F. L. (2003). *Python language reference manual*. Network Theory.

[30] Oliphant, T. E. (2007). "Python for scientific computing". *Computing in Science & Engineering* 9.3, pp. 10–20.

[31] Hunter, J. D. (2007). "Matplotlib: A 2D graphics environment". *Computing in Science & Engineering* 9.3, pp. 0090–95.

[32] Talbot, H. (2000). "wxPython, a GUI Toolkit". *Linux Journal* 2000.74es, p. 5.

[33] Tseng, Y.-H., Uetrecht, C., Heck, A. J., and Peng, W.-P. (2011). "Interpreting the charge state assignment in electrospray mass spectra of bioparticles". *Analytical Chemistry* 83.6, pp. 1960–1968.

[34] Marx, M. L. and Larsen, R. J. (2006). *Introduction to mathematical statistics and its applications*. Pearson/Prentice Hall.

[35] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical recipes in C*. Cambridge University Press.

[36] Marquardt, D. W. (1963). "An algorithm for least-squares estimation of nonlinear parameters". *Journal of the Society for Industrial & Applied Mathematics* 11.2, pp. 431–441.

[37] Scarff, C. A., Thalassinos, K., Hilton, G. R., and Scrivens, J. H. (2008). "Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements". *Rapid Communications in Mass Spectrometry* 22.20, pp. 3297–3304.

[38] Smith, D. P., Knapman, T. W., Campuzano, I., Malham, R. W., Berryman, J. T., Radforda, S. E., and Ashcroft, A. E. (2009). "Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies". *European Journal of Mass Spectrometry* 12.13, p. 13.

[39] Bush, M. F., Hall, Z., Giles, K., Hoyes, J., Robinson, C. V., and Ruotolo, B. T. (2010). "Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology". *Analytical Chemistry* 82.22, pp. 9557–9565.

[40] Hilton, G. R., Thalassinos, K., Grabenauer, M., Sanghera, N., Slade, S. E., Wyttenbach, T., Robinson, P. J., Pinheiro, T. J., Bowers, M. T., and Scrivens, J. H. (2010). "Structural analysis of prion proteins by means of drift cell and traveling wave ion mobility mass spectrometry". *Journal of the American Society for Mass Spectrometry* 21.5, pp. 845–854.

[41] Zhong, Y., Hyung, S.-J., and Ruotolo, B. T. (2011). "Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions". *Analyst* 136.17, pp. 3534–3541.

[42] Nyon, M. P., Segu, L., Cabrita, L. D., Lévy, G. R., Kirkpatrick, J., Roussel, B. D., Patschull, A. O., Barrett, T. E., Ekeowa, U. I., Kerr, R., Waudby, C. A., Kalsheker, N., Thalassinos, K., Lomas, D. A., Christodoulou, J., and Gooptu, B. (2012). "Structural dynamics as-

sociated with intermediate formation in an archetypal conformational disease". *Structure* 20.3, pp. 504–512.

[43] Leary, J. A., Schenauer, M. R., Stefanescu, R., Andaya, A., Ruotolo, B. T., Robinson, C. V., Thalassinos, K., Scrivens, J. H., Sokabe, M., and Hershey, J. W. (2009). "Methodology for measuring conformation of solvent-disrupted protein subunits using T-WAVE ion mobility MS: an investigation into eukaryotic initiation factors". *Journal of the American Society for Mass Spectrometry* 20.9, pp. 1699–1706.

[44] Freeke, J., Robinson, C. V., and Ruotolo, B. T. (2010). "Residual counter ions can stabilise a large protein complex in the gas phase". *International Journal of Mass Spectrometry* 298.1, pp. 91–98.

[45] Hopper, J. T. and Oldham, N. J. (2009). "Collision induced unfolding of protein ions in the gas phase studied by ion mobility-mass spectrometry: the effect of ligand binding on conformational stability". *Journal of the American Society for Mass Spectrometry* 20.10, pp. 1851–1858.

[46] Hyung, S.-J., Robinson, C. V., and Ruotolo, B. T. (2009). "Gas-phase unfolding and disassembly reveals stability differences in ligand-bound multiprotein complexes". *Chemistry & Biology* 16.4, pp. 382–390.

[47] Warelow, T. P., Oke, M., Schoepp-Cothenet, B., Dahl, J. U., Bruselat, N., Sivalingam, G. N., Leimkühler, S., Thalassinos, K., Kappler, U., Naismith, J. H., and Santini, J. M. (2013). "The respiratory arsenite oxidase: structure and the role of residues surrounding the Rieske cluster". *PloS One* 8.8, e72535.

4.5 Appendix

4.5.1 Example maximum entropy (MaxEnt) spectrum

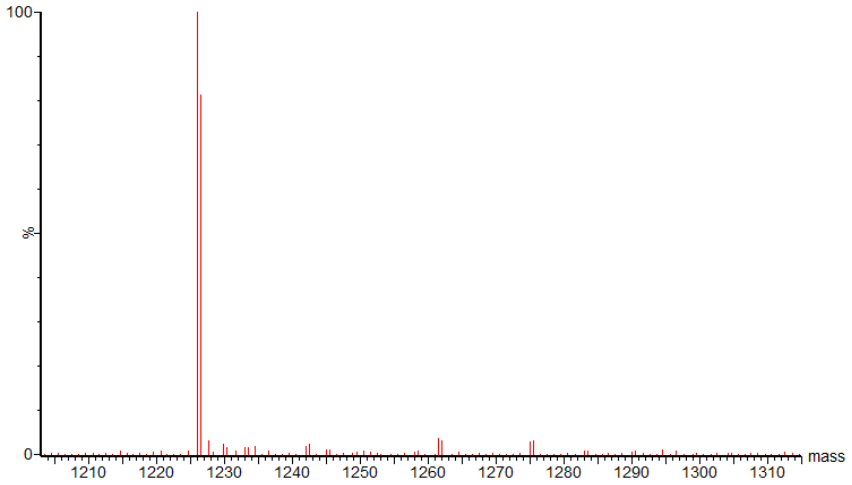

Figure 4.15: Maximum entropy spectrum produced by the MaxEnt 3 function of Waters MassLynx software. The most abundant molecular species is a 1,226 Da peptide.

4.5.2 Simplifying calibration equations

Determining corrected arrival time t'_d

The following equation is the same as Equation 3.7, the parameters other than the corrected arrival time are; the experimental arrival time (t_d), the portion of the arrival time which is independent of m/z ($t_{independent}$) and the m/z dependent time contribution ($t_{dependent}$), such as time spent in the ToF.

$$t'_d = t_d - t_{independent} - t_{dependent} \tag{4.12}$$

Substitute in $t_{independent}$ from Equation 3.8 and simplify. The new parameter V_w is T-wave velocity measured in m/s.

$$t'_d = t_d - (61 + 31)\left(0.01 \cdot \frac{300}{V_w}\right) - t_{dependent} \tag{4.13}$$

$$t'_d = t_d - 92 \cdot \frac{3\cancel{00}}{1\cancel{00} \cdot V_w} - t_{dependent} \tag{4.14}$$

$$t'_d = t_d - \frac{276}{V_w} - t_{dependent} \tag{4.15}$$

Substitute in $t_{dependent}$ using Equation 3.9 and simplify. ξ is the mass-to-charge ratio of the ion.

$$t'_d = t_d - \frac{276}{V_w} - \sqrt{\frac{\xi}{1000}} \cdot (0.044 + 0.041) \tag{4.16}$$

$$t'_d = t_d - \frac{276}{V_w} - 0.085 \cdot \sqrt{\frac{\xi}{1000}} \tag{4.17}$$

Determining collision cross section (Ω)

The following equation is the same as Equation 3.12 which calculates the collision cross section using the powerfit parameters (A and B), the reduced mass (μ), charge state (z) and corrected arrival time.

$$\Omega = A \cdot \mathrm{t}_d'^B \cdot z \cdot \sqrt{\frac{1}{\mu}} \tag{4.18}$$

Substitute in the corrected arrival time (t_d') using the new Equation 4.17 and simplify.

$$\Omega = A \cdot z \cdot \sqrt{\frac{1}{\mu}} \cdot \left(\mathrm{t}_d - \frac{276}{V_w} - 0.085 \cdot \sqrt{\frac{\xi}{1000}} \right)^B \tag{4.19}$$

$$\Omega = \frac{A \cdot z}{\sqrt{\mu}} \cdot \left(\mathrm{t}_d - \frac{276}{V_w} - 0.085 \cdot \sqrt{\frac{\xi}{1000}} \right)^B \tag{4.20}$$

4.5.3 Amphitrite graphical user interfaces

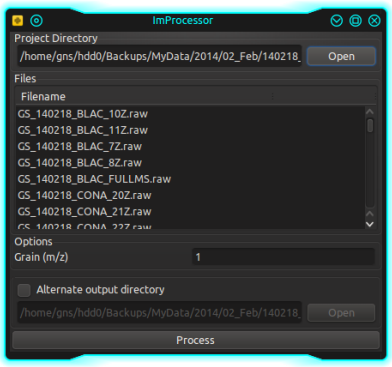

Figure 4.16: ImProcessorGui from Amphitrite. This program allows users to batch convert MassLynx ion mobility files into Amphitrite data files (`.a`). The user selects the directory containing data files (usually the data from a day's experiment), they can then select which files are to be converted. The conversion process currently only works in Windows as the MassLynx conversion library only supports this operating system.

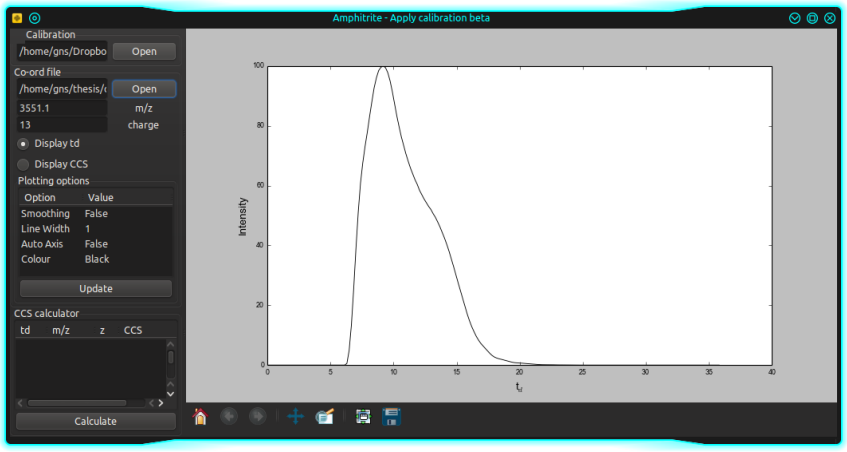

Figure 4.17: ApplyCalibrationGui from Amphitrite. Uses data from "Copy Spectrum List" of an ATD in MassLynx. Allows for the display of ATDs and the calibration to CCS using an Amphitrite calibration file.

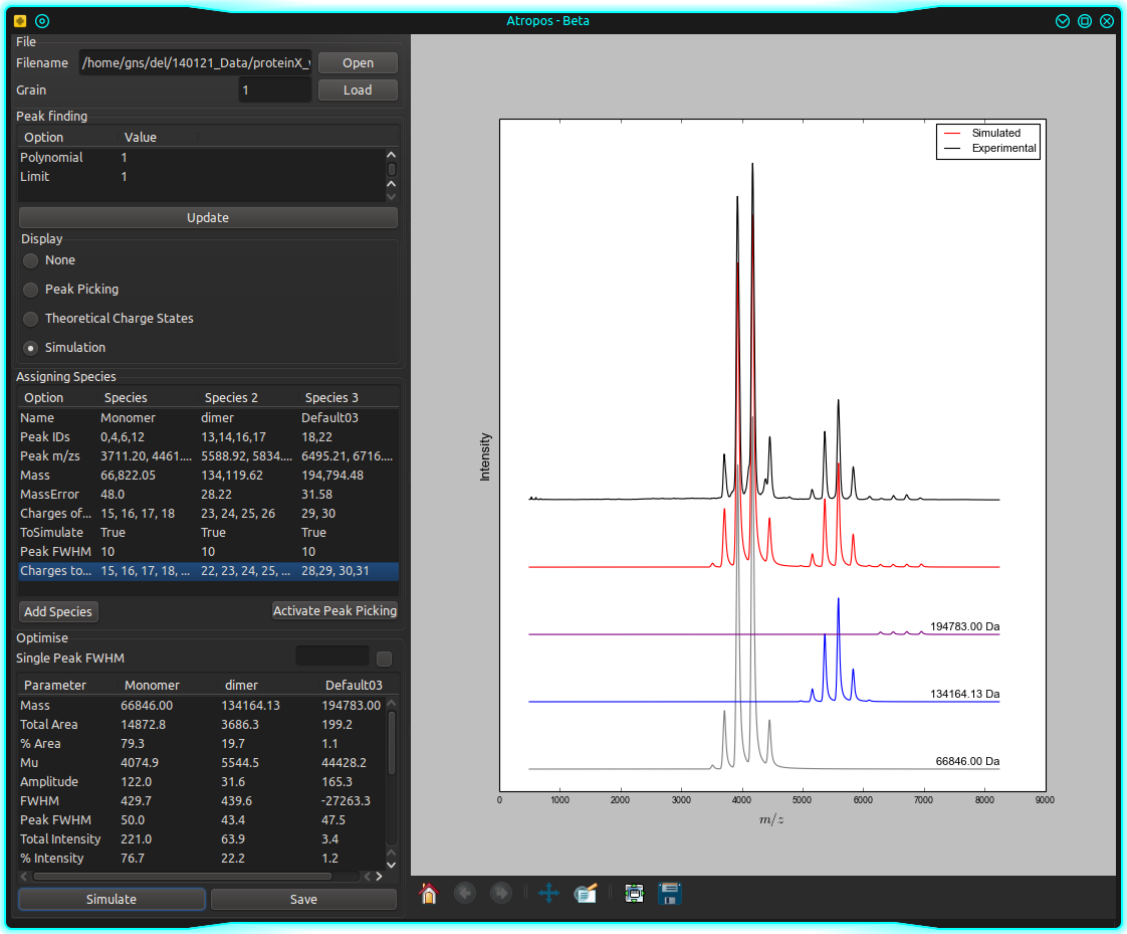

Figure 4.18: AtroposGui from Amphitrite. This program allows the user to deconvolute mass spectra and export the deconvolution parameters as a Amphitrite mass spectrum fit (`.afit`) file to be used with other Amphitrite programs. The input data can be from "Copy Spectrum List" of a mass spectrum in MassLynx, or as the output data from Amphitrite's ImProcessorGui.

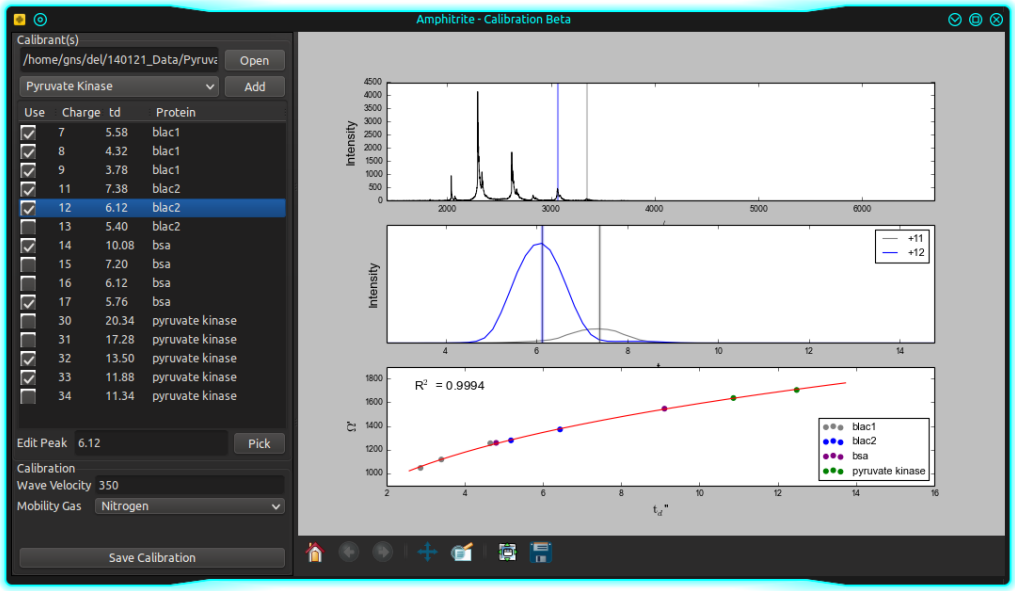

Figure 4.19: CalibrationGui from Amphitrite. CalibrationGui allows the user to automatically extract the ATD for each charge state of a calibrant protein, and automatically plot a calibration curve. Individual charge states can be removed if of low quality, the position of the ATD peak can also be moved. The calibration can then be exported as an Amphitrite calibration file (`.acal`).

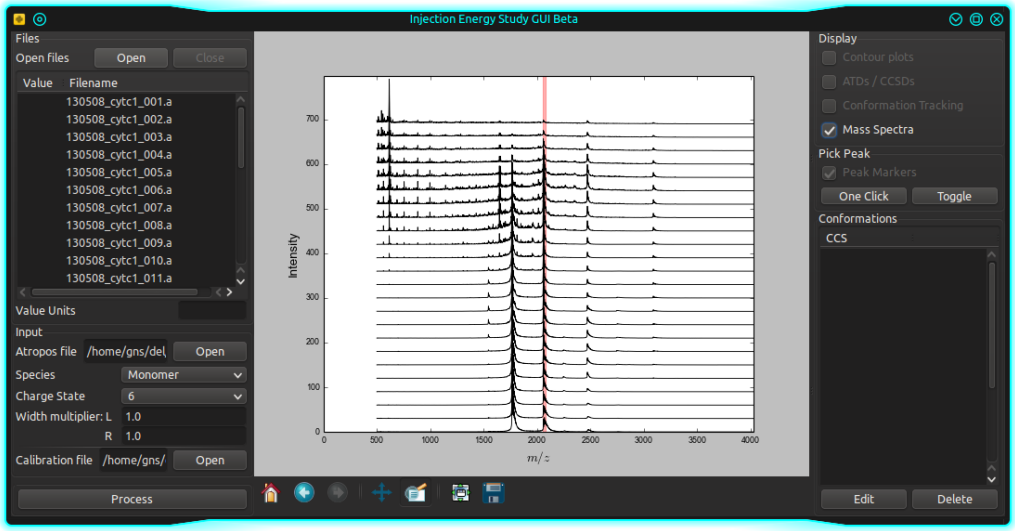

Figure 4.20: IesGui from Amphitrite. This GUI allows users to analyse multiple acquisitions of the same protein simultaneously and was designed to analyse collision induced unfolding data. After multiple Amphitritre data files are loaded, with an Amphitrite fit file, a particular charge state can be extracted, thereby showing the ATDs, and unfolding contour plots. These can additionally be calibrated to show the data in terms of CCS instead of arrival time. The program can also monitor abundance of conformational families using peak heights in ATDs/CCSDs.

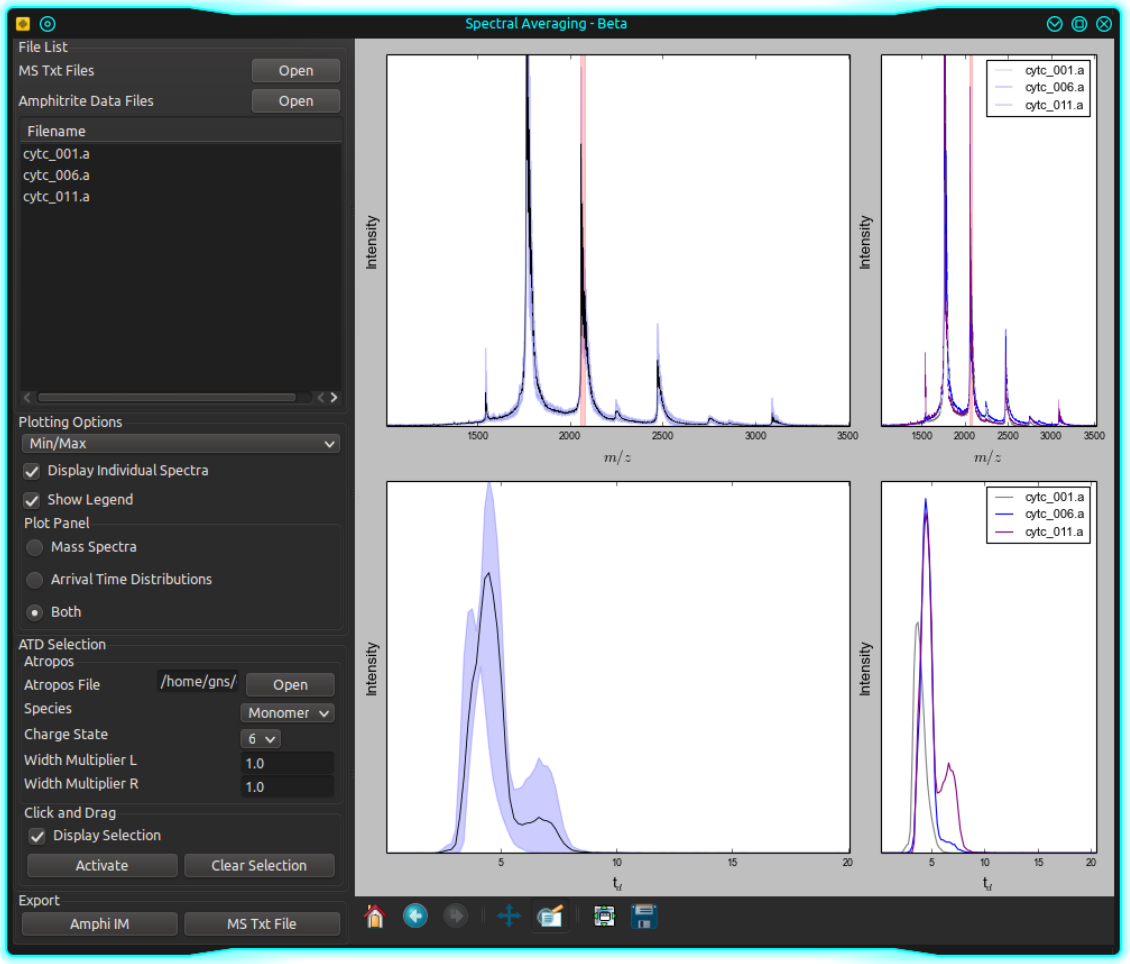

Figure 4.21: SpectralAveragingGui from Amphitrite. Spectral averaging allows users to produce a representative spectrum (top) and/or ATD (bottom) from replicated acquisitions. The GUI allows users to do this with MS, ATD and CCSD data. The left plots show the average in black, with the range between minimum to maximum values across the three spectra shaded in blue. On the right the three original mass spectra and ATDs are shown. data are taken from 3 voltages of CIU to make the effect clear. The user can then export the data either as an Amphitrite data file or a text file.

Chapter 5

Challenger

In the previous chapter, basic collision induced unfolding (CIU) data analysis was introduced. This chapter introduces several new methods for analysing these data, which aim to be more accurate and quantitative than existing methods.

5.1 Introduction

5.1.1 Gas-phase unfolding of proteins

Many studies have shown that with careful control of instrument parameters, protein collision cross section (CCS) values obtained using ion mobility mass spectrometry (IM-MS) are closely related to their native solution structure [1–3]. It was also determined that during IM-MS higher charge states represent unfolded species, increasing the reported CCS. It was found that to achieve the most native-like conformation the lowest charge state should be used [4, 5]. This raises the question, would it be useful to intentionally unfold proteins in the gas-phase?

Proteins with higher charge states are likely to be more unfolded due to the coulombic repulsion between the adducted protons causing disruption to the protein structure. Additionally, the kinetic energy of ions by electric

fields in the mass spectrometer is proportional to the number of charges on the protein. The additional velocity results in more violent collisions with inert gas molecules, causing ion heating, due to friction, and subsequently causes protein unfolding. Analysing protein unfolding using ion charge state is suboptimal as the charges on the ion are integers, so transient conformations could be missed. Another problem is that coulombic repulsion does not relate to a process that happens in solution, whereas collisional heating is similar to the effect of solution temperature as the unfolding is caused by gradual increasing the internal energy of the ion, as heat, through many collisions.

The early experiments using drift tube instruments were capable of protein unfolding IM-MS experiments. Quadrupole isolated ions were accelerated into the drift cell, where collisions with the mobility gas caused ion heating and unfolding. The ions quickly decelerate and ion mobility data can be acquired. These experiments are known as injection energy studies [6].

Injection energy studies were used by the Clemmer group [7] to show that disulphide bonds stabilise protein conformation in the gas-phase as well as in solution and the observation indicated that alterations to solution structure stability could also affect gas-phase stability.

Another use of injection studies by the Bowers group was to determine the oligomeric states of ions that had been quadrupole isolated [8]. When analysing oligomers of Aβ42, it was difficult to determine the oligomeric size as, for example, a dimer with twice the number of charges as a monomer will have the same m/z value. This problem was compounded by the low signal intensity achieved when analysing these oligomers. By using increments in injection energy, they caused the oligomer to dissociate and determined the oligomeric state of the original ion and product ions by the IM-MS drift time [8, 9].

Another technique that causes the collisional heating of ions is collision induced dissociation (CID) [10, 11]. The instrumentation used is a dedicated collision cell, where ions are accelerated by an electric field and collisions with inert gas cause the increase in the internal energy of the ions. The Waters Synapt travelling wave ion mobility mass spectrometer has a collision cell

after a quadrupole mass analyser and before the IM separation cell and time-of-flight (ToF) mass analyser. The instrumentation allows for the initial mass selection of an ion charge state, followed by ion mobility separation and mass determination using the ToF.

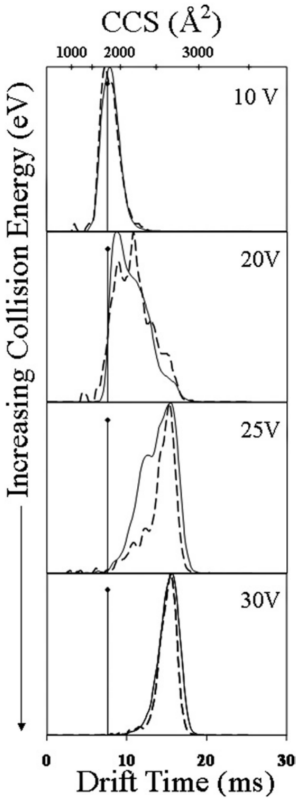

Figure 5.1: Gas-phase unfolding of apo (dashed trace) and holo (solid trace) myoglobin. Vertical line indicates the CCS of holo myoglobin as calculated from an X-ray crystallography structure. Reproduced from [12].

An important use for gas-phase unfolding is to investigate the stability of a protein in the presence and absence of a ligand. Hopper and Oldham investigated the stability of apo and holo myoglobin by measuring changes in arrival time distributions caused by increasing the collision energy using TWIM-MS [12]. The results are shown in Figure 5.1. At low collision energies both forms of the protein have similar CCS values, however as the collision energy is increased, the apo form unfolds more readily. Solution-phase experiments have found holo myoglobin to be more stable than the apo form [13], and so the TWIM-MS data shows that this characteristic is maintained in the gas-phase.

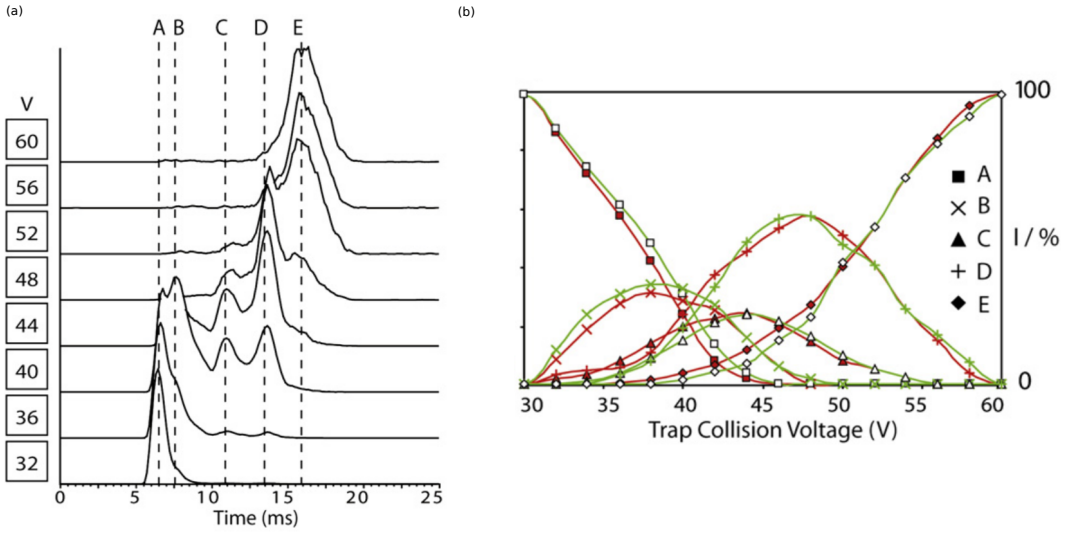

Figure 5.2: Gas-phase unfolding and conformational abundance tracking of wild type and L55P mutant tetrameric transthyretin (TTR). (a) Arrival time distributions of the +15 charge state of wild type TTR at increasing collision energies. Individual conformations are labelled A-E. (b) Conformational abundance tracking of each conformation A-E using the ATD peak intensity for wild type (filled) and L55P mutant (open) TTR. Conformation abundance is reported as the proportion of the summed peak intensities for all conformations per ATD. Figure reproduced from [14].

The data reported for gas-phase unfolding studies have shown that protein and protein complexes do not gradually unfold in a linear manner, rather unfolding occurs through a series of distinct transient conformational families [4, 14–18].

The CCS values of each individual ion (of a particular analyte), when aggregated, form a Gaussian distribution. The reason for this is that several structural elements of a protein will be interconverting between different positions, and the conformational family (also referred to as a conformation) is a mixture of all of these small movements. With distinct conformations, a non-interconverting structural change has taken place, and this once again has a Gaussian distribution of CCS values as the smaller vibrational movements continue to occur around a different mean CCS value.

The static mean of a distinct conformational population means that the abundance of individual conformational families in an unfolding experiment can be determined by monitoring the peak height at the mean and an example

is shown in Figure 5.2 [14]. Ion mobility data are of low resolution and so conformation peaks often overlap, as seen with conformations A and B in Figure 5.2(a). This leads to distortion in the reported abundances when analysing ATDs using conformational peak heights. In some studies, researchers have used statistical software such as Fityk [19] to fit Gaussian distributions to single ATDs in order to give a better representation of the abundance of conformations [20, 21].

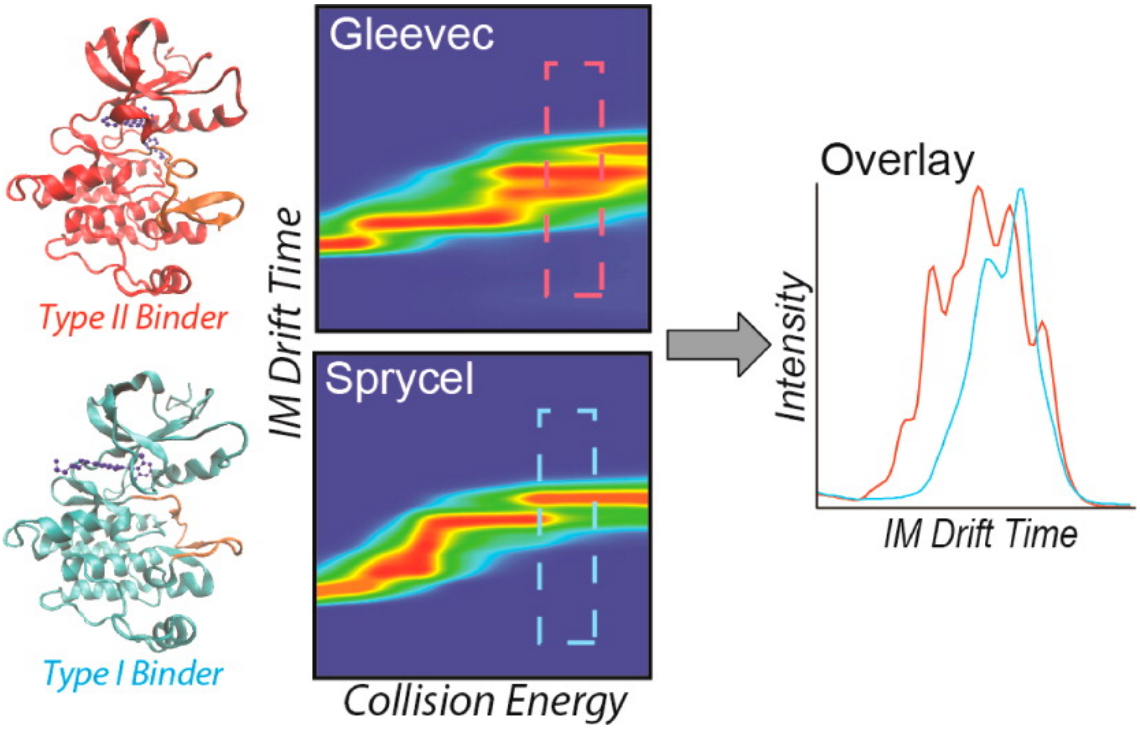

Figure 5.3: Demonstration of CIU fingerprint analysis, applied to the protein kinase domain of Bcr-Abl. (left) Protein structure diagrams showing the binding position of type I and type II inhibitors. Gas-phase unfolding data, of the complexes from each inhibitor class, are represented as CIU fingerprints (centre). ATDs are extracted from the data in the dashed area of the CIU fingerprints (right). This shows that there are clear differences between the way the protein unfolds when bound to type I or type II inhibitors. Figure reproduced from [17]

A new method for analysing gas-phase protein unfolding data was recently developed called collision induced unfolding (CIU) fingerprinting, which gives a visual representation of the pattern of unfolding. The method involves extracting ATDs for a particular ion across several acquisitions at voltage increments. These these sets of two-dimensional data are stacked with the voltage increment along the x axis, the arrival time along the y axis and the

ion intensity as the z axis to give a three-dimensional heatmap, an example of which is shown in Figure 5.3.

Using CIU fingerprinting, it was demonstrated that although binding type I or type II kinase inhibitors of the protein kinase domain of Bcr-Abl did not substantially alter the conformation at low collision energies, gas-phase unfolding the protein and analysing the data with CIU fingerprints showed different unfolding patterns for each of the types of ligands. These different unfolding patterns could then be used to class putative inhibitors as having a type I or II binding mechanism, without the need for obtaining X-ray crystallography structures for each complex (Figure 5.3) [17].

5.1.2 Aims

The introduction of commercially available instrumentation for IM-MS analysis [16, 22] has led to an expansion in the number of laboratories that have access to such technologies. A consequence of new types of data acquisition methods, however, is the need for development of new computational methods that are able to cope with the complexity and extract the maximum information from such new data, something that has been lacking for the IM-MS field. To ameliorate this, we developed the software package Amphitrite [23] introduced in Chapter 4. It can be used to deconvolute the MS data and use this information to then automatically extract ATDs from the raw data files and automatically apply a collision cross section (CCS) calibration to them. Amphitrite introduced some basic methods of analysis for CIU data, including the ability to assess the abundance of conformations using peak heights.

The open source codebase for Amphitrite makes it easy to extract ATDs from MassLynx data files in an automated fashion, and then manipulate the data programmatically. This development led us to investigate the shortcomings of current methods of data analysis used with gas-phase unfolding data.

The first problem is the quantitative representation of large unfolding datasets. After a certain number of voltage increments, representing the data

as stacked arrival time distributions is unfeasible. The alternative CIU fingerprint allows for a useful visual representation of the unfolding process, however, the data are not quantitative and so statistical or automated comparisons between analytes would not be possible. In order to achieve these characteristics, methods of numerically summarising ATDs that go beyond using peak tops were investigated and are presented here.

The second issue was the representation of conformational abundances. Tracking the abundance of transient partially folded species during unfolding experiments can impart important information when comparing two analytes. The peak height method is susceptible to overlapping peaks and so a deconvolution approach is proposed which utilises a purpose built GPU (graphics processing unit) accelerated genetic algorithm.

The frequent focus of ion mobility unfolding experiments on ligand interactions suggests the possibility of using the technique for testing stability in high-throughput drug screening. A method and tool to automate unfolding experiments on the Waters Synapt instrument are presented here. In combination with the ability for automated ATD extraction afforded by Amphitrite, and automatic sample changing using tools like the Advion Nanomate, this work could increase the viability of the technique in drug screening experiments.

5.2 Methods

5.2.1 Sample sources

Lysozyme from hen egg white, myoglobin from equine heart, β-lactoglobulin from bovine milk, concanavalin A from *Canavalia ensiformis* and bovine serum albumin (BSA) were purchased from Sigma Aldrich (St. Louis, MO).

5.2.2 Mass spectrometry sample preparation

The proteins were analysed in 200 mM ammonium acetate which was purchased from Sigma Aldrich (St. Louis, MO). Buffer exchange was carried out using BioRad (Hercules, CA) BioSpin 6 columns, with additional concentration and dilution steps using Amicron Ultra 0.5ml centrifuge filters (Millipore UK Ltd, Watford, UK). The concentration of these samples were monitored using a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA), samples were analysed after diluting to 10 μM.

5.2.3 IM-MS procedures

The experiments were performed on a Synapt HDMS mass spectrometer (Waters Corp., Manchester, UK) [16]. The instrument was mass calibrated using 30 μM caesium iodide (Sigma Aldrich, St. Louis, MO) dissolved in 250 μM ammonium acetate. 2.5 - 4.5 μl aliquots of sample were delivered to the mass spectrometer using gold coated borosilicate capillaries, which were prepared *in house* [24].

Calibration curves were calculated using BSA and bradykinin [25] acquired using the same ion mobility parameters as the samples. The data was extracted and calibrations applied using the Amphitrite software package [23, 26].

Ion mobility experiments were carried out after quadrupole isolation of a particular charge state for each analyte. This allows for the analysis of dissociation products of the CID process without interference from different charge states which receive different energies at a given voltage. Additionally the T-wave ion mobility device can exhibit reduced separation due to charge state saturation, by using quadrupole isolation the ions entering the IM cell are limited to those that are to be analysed during the experiment.

Ion mobility data was acquired after quadrupole isolation of a particular charge state (+7 lysozyme, +8 myoglobin and +8 β-lactoglobulin). The instrument settings for the unfolding experiments are shown in Table 5.1.

Setting	Value
Capillary Voltage	1.1 kV
Sampling Cone	20 V
Extraction Cone	1 V
Source Temperature	40 °C
Trap Collision Energy	Variable
Transfer Collision Energy	5 V
Trap Pressure	5 $m$bar
Mass Range	1,000-8,000 m/z
Bias Voltage	20 V
Backing Pressure	0.24 $m$bar
IM Wave Height	9 V
IM Wave Velocity	350 $m \cdot s^{-1}$
Transfer Wave Height	8 V
Transfer Wave Velocity	150 $m \cdot s^{-1}$
IMS pressure	5.20x10^{-1} $m$bar
LM/HM Resolution	5/15

Table 5.1: IM-MS settings table for unfolding experiments on model proteins.

5.2.4 Genetic algorithms

Genetic algorithms are a class of artificial intelligence optimisation algorithms, which mimic evolution in order to find an optimal solution. At the start of the process, a randomly generated population of individuals (also known as chromosomes) is created. An individual contains all the parameters that are to be optimised, which are known as genes. For this algorithm, individuals can contain values for calculating a Gaussian distribution such as amplitude and full width half maximum (FWHM).

The next stage is to select the better optimised (higher fitness) individuals for two new populations. This is carried out using tournament selection, for each individual in each population n individuals are randomly selected from the previous population and the fittest individual is used. The fitness of an individual is determined by simulating the Gaussian distributions that they describe, summing them together and calculating the absolute difference to each of the experimental data points. The additive inverse of the square of the

summed error vector is used as the fitness, resulting in a maximum fitness of zero. The two populations are then merged into one through recombination, an individual is taken from each population (parents) and a new individual is created by randomly selecting genes. The last step is to introduce mutations into the new population. A mutation rate is given as a probability. This is applied to all genes and individuals, and where mutation occurs a new value is randomly generated, within a preset range of values.

At this stage, the fittest individual from the population is determined and if its error is sufficiently low, the algorithm terminates and the fittest individual is returned as the solution. If the error is higher, the population undergoes another round of evolution, starting from the tournament selection step, however during this round the fittest individual is not changed in a process called elitism. If the algorithm does not find the solution after a certain number of generations, the algorithm will terminate, returning the fittest individual as its solution [27].

5.2.5 Software development

Three implementations of the Challenger algorithm were developed using different technologies. The first was developed in the Python programming language [28] and a Python genetic algorithm library, PyEvolve [29]. The second version was developed in the C programming language [30]. The final version, which is used throughout this chapter, was written in Cuda C [31]. The final Cuda C version has been tested on GNU/Linux (64 bit) and Microsoft Windows 7 (64 bit).

The Python programming language [28], in conjunction with Python modules NumPy, SciPy [32, 33], Matplotlib [34], and Amphitrite [23], handles the processing and displaying the results generated by the genetic algorithm.

Processing times quoted are for a GNU/Linux desktop computer with an Intel i5 3570K overclocked to 4.4 GHz, 16 GB RAM and an nVidia GeForce GTX 660 Ti graphics processing unit (GPU).

5.3 Results and discussion

5.3.1 Automation of unfolding experiments

Increasing the collision energy increases the amount of kinetic energy experienced by the ion traversing the trap cell. Whilst the ion is crossing the trap cell, it experiences collisions with inert gas molecules, increasing the internal energy of the molecule and causing it to unfold. Monitoring the resultant unfolding by means of IM-MS have been important experiments for the field. They have also gone by different names including injection energy studies [5] and collision induced unfolding [35]. The procedure involves acquiring short (1-2 minute) sets of data, altering the collision energy and reacquiring, repeatedly. It seemed that this procedure would be well suited to automation.

A program was developed to generate tune files containing instrument settings that were compatible with the Waters Synapt G1, and it should be straightforward to adapt for use with newer Waters Synapt instruments. A screenshot of the program's graphical user interface is shown in Figure 5.4. Using this software, sets of tune files can be generated with ascending trap cell collision energies for use in the experiments described here. Additionally other parameters can be altered, for instance the sampling cone collision voltage could be ramped if the user needed to use quadrupole mass selection after the unfolding event.

The instrument's software (MassLynx 4.1) can be set to be able to trigger and cease data acquisitions. This allows the user to create a sample list file, and enter the different tune files to be used for each acquisition. This process has to be done manually so the program outputs tune files with names in the format `ipr_xxx.ipr` where `xxx` is the voltage increment index. For example when acquiring an unfolding experiment which starts with a trap voltage of 5 V and increases in 5 V increments, then the first tune file will be labelled `ipr_001.ipr` (5 V), and the second `ipr_002.ipr` (10 V). This means that the manually entered sample list can be used for many experiments with different sets of automatically generated tune files.

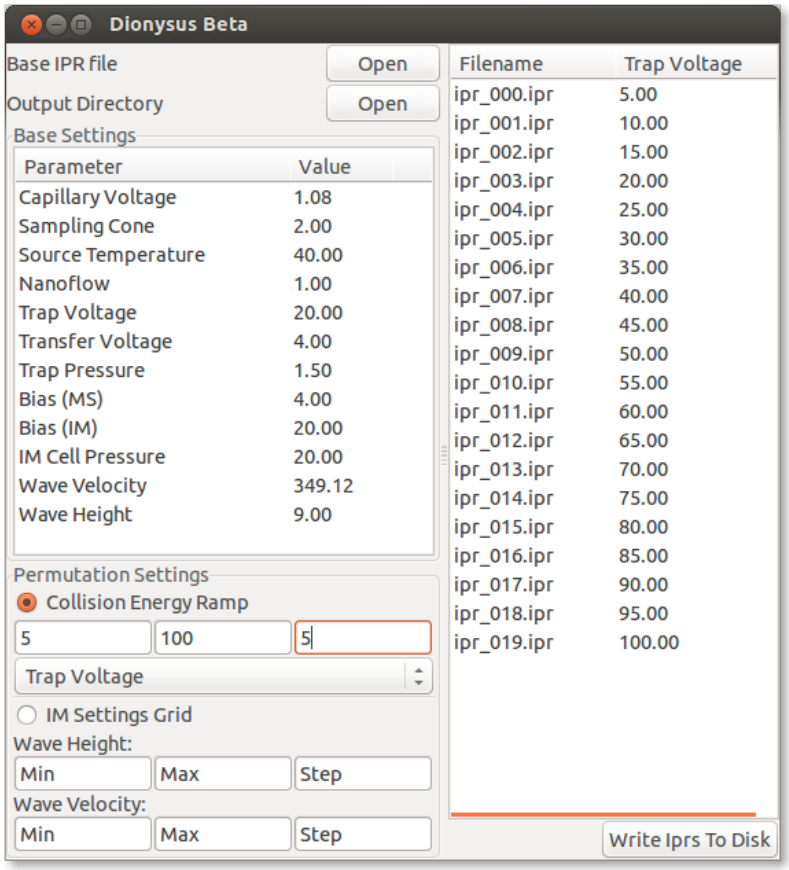

Figure 5.4: GUI for automatically generating mass spectrometer settings files (.ipr) for the Waters Synapt.

When using this method the dead time between acquisitions is consistently under 2 seconds, this contrasts to a user based method which would be slower at its fastest, and would not have the same level of consistency.

This automation procedure also facilitates overnight experiments, very fine grained unfolding analyses could be left to run for several hours making use of time when the instrument is not being used and the increased amount of data reduces the chance of missing conformations which are only present within small collision energy windows.

Once these data have been acquired, the arrival time distributions need to be extracted from the numerous raw data files. Using MassLynx to accomplish this extraction is a multi-step process for each file, including an important operation where the m/z range of the peak in the mass spectrum is selected by the user. Using the Amphitrite software package [23] all data files can be

processed and extracted in an automated fashion. The *m/z* range used can either be entered by the user or can be calculated using a mass spectrum fit generated by the program, thereby removing variability in extraction of data.

The automation of data acquisition would allow for 24 hour operation for high-throughput experiments when used alongside a system such as the Advion Nanomate to switch between samples.

5.3.2 Summarising IM-MS unfolding data

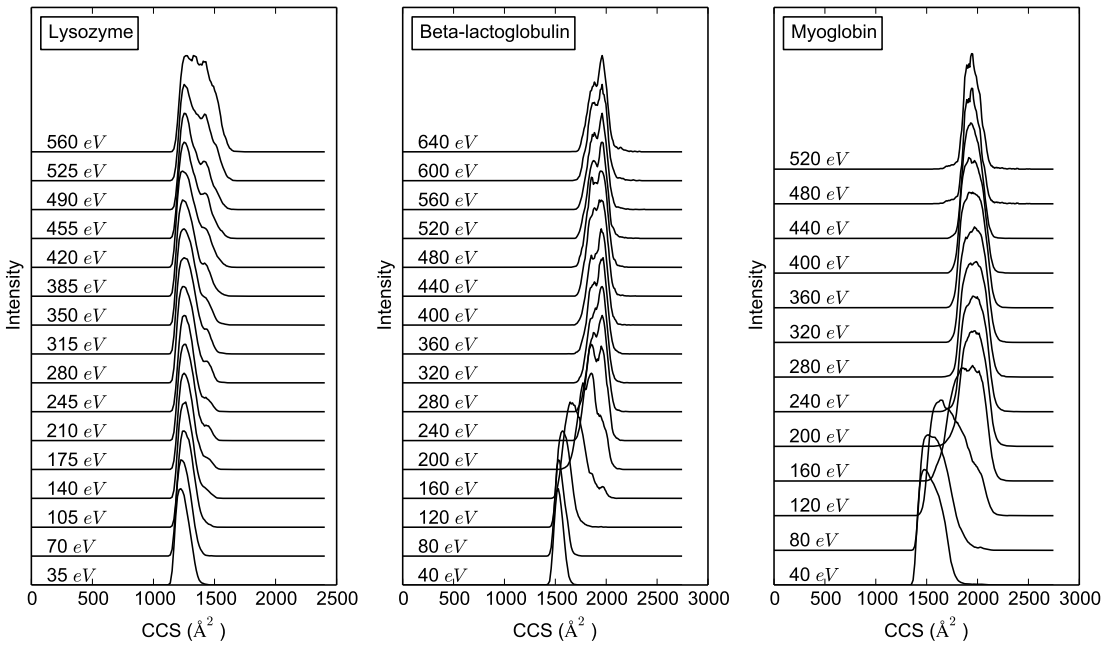

Figure 5.5: Arrival time distributions of the gas-phase unfolding of lysozyme, β-lactoglobulin and myoglobin.

Unfolding curves

The large amount of data produced by collision energy ramps can be difficult to represent graphically. In some cases in the literature 3-5 ATDs are displayed by vertically stacking them [12], this allows a figure to show the key points of each distribution clearly. However, when increasing the number of distributions represented in this way to more than 10, it becomes difficult to see important features (see Figure 5.5). Additionally it would be helpful in a drug screening environment to have a quick overview of whether a particular

ligand changes the conformational flexibility or stability of the target protein. This way ligands that do not have the desired effect can be discarded without thorough investigation of the unfolding process.

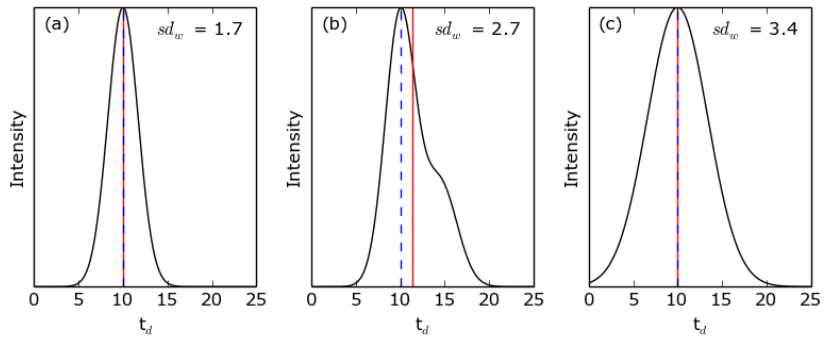

Figure 5.6: Dummy data representing different arrival time distributions which would have the same arrival time based on peak top value; (a) sharp distribution,low variation in protein conformation, (b) compound distribution, consisting of two conformations but where one is not a peak shoulder, (c) wide single distribution, representing large structural flexibility without distinct conformational families. The t_d corresponding to the peak apex is indicated by a dashed blue line,the red line indicates the weighted average t_d value, and the values for weighted standard deviation and area under curve are shown in the top right of each panel.

Figure 5.6 shows three synthetic arrival time distributions, in all three distributions the maximum intensity is found at an arrival time of 10 ms. Though the distributions are quite different, peak top analysis would represent them to be the same.

Distribution (a) is comprised of a single narrow conformational family and (b) has an additional conformation at a later arrival time. In order to summarise the arrival time value of these distributions, additional conformations after the most abundant would have to be included in the calculation.

To determine a number which represents the central arrival time of an ion, aggregated over all potential conformations present, the weighted mean was calculated (Equation 5.1, where I is the intensity, t is arrival time and n is the number of data points in the arrival time axis). This metric is indiscriminate of individual conformations and so can accurately summarise the bulk conformation.

$$\overline{t_w} = \frac{\Sigma_{i=1}^{n} I_i t_i}{\Sigma_{i=1}^{n} I_i} \tag{5.1}$$

For analysing the protein standards, the arrival time values are converted to collision cross section (CCS) before the calculation, additionally the voltage was converted to laboratory voltage, eV (charge state multiplied by voltage). This allows different proteins, of different charge states to be directly compared and the result is shown in Figure 5.7(A).

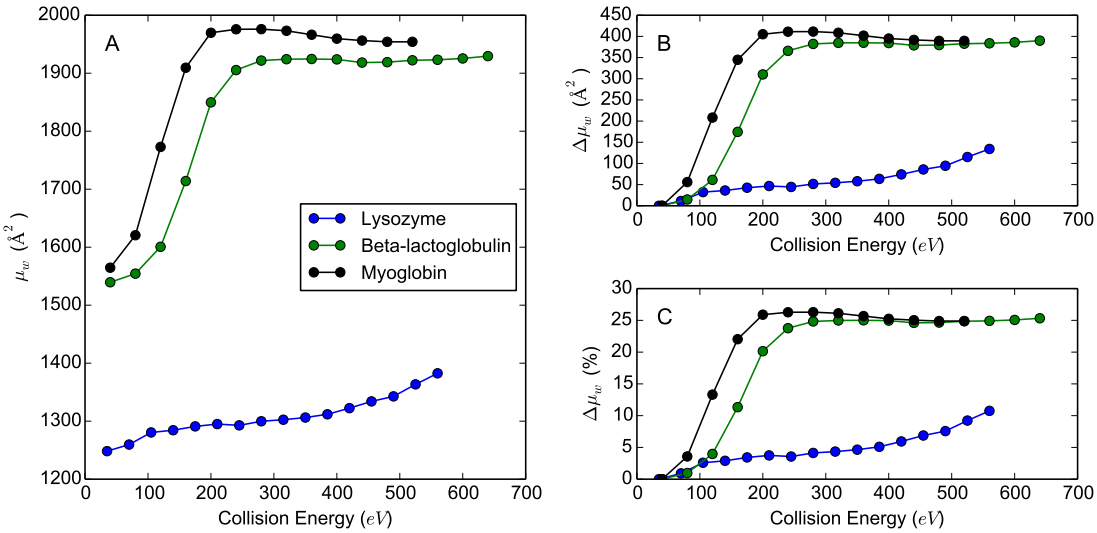

Figure 5.7: Unfolding curves for model proteins monitoring; collision cross section (CCS) (A), change in cross section in comparison to the lowest collsion energy acquisition, monitored as, absolute CCS (B) and percentage change in CCS (C).

β-lactoglobulin and myoglobin are larger proteins (18.4 kDa and 17.6 kDa respectively), and so, as expected the smaller protein, lysozyme (14.2 kDa) starts with a smaller CCS (Figure 5.7(A)). Lysozyme is highly stable in part due to its structure being maintained by four disulphide bonds [7, 36]. This explains the much lower increase in cross section, in comparison to the other proteins, as the collision energy is increased. The structure of β-lactoglobulin is maintained by two disulphide bonds [37] as opposed to none for myoglobin, potentially explaining the difference in gas phase stability between them. To aid in the visualisation, a similar graph is shown in Figure 5.7(B), where the change in cross section from the lowest voltage aquisition is shown.

The experiment was carried out on the holo version of myoglobin. Due to the two stage mass analysis of the instrument (m/z window isolation using the quadrupole, followed by ToF analysis), the apo form can be excluded from the final ATD analysed. It has been previously shown that the protein is more stable in holo form [13].

After the CCS of myoglobin reaches its highest value, the CCS starts to reduce. This effect is likely due to the haeme group dissociating from the most extended conformational family, causing a m/z shift out of the analysed range and consequently we see an over representation of the more compact conformations (Figure 5.5 and Appendix 5.20).

Conformational variability curves

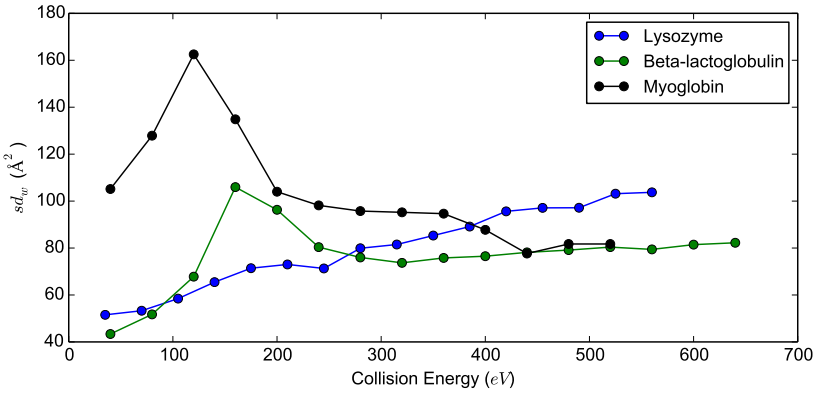

Figure 5.8: Monitoring the conformational variability of three model proteins during gas-phase unfolding.

In addition to the mean CCS, analysing the bulk sample can also produce information regarding conformational flexibility. This is demonstrated by the comparison of distributions in Figure 5.6(a) and (c). Distribution (c) is much wider, which would indicate a greater conformational flexibility. In this case, the peak top analysis would not be misrepresenting the data, as both distributions are symmetrical around the same centre, though the information regarding conformational flexibility would be lost.

The weighted standard deviation (Equation 5.2) can be compared between multiple arrival time distributions directly. This produces a value which is representative of the conformational flexibility displayed in the distribution.

The calculated standard deviation values are indicated in Figure 5.6, and as expected distribution (a) has the lowest value, and distribution (c) being higher than (b).

$$sd_w = \sqrt{\frac{\Sigma_{i=1}^{n} I_i(t_i - \overline{t_w})^2}{\frac{m-1}{m}\Sigma_{i=1}^{n} I_i}} \tag{5.2}$$

Applying this calculation to the unfolding data for the model proteins yields Figure 5.8. Regions with higher standard deviation values are indicative of conformational change; the protein likely exists in several transient conformations increasing the spread of the ATD. This behaviour is observed in myoglobin at 100-200 V, following this there is a two-stage decline. This phenomenon is observed when the protein stabilises a particular conformational family, and the second decline is likely caused by the dissociation of a more extended conformation (see Appendix 5.20).

The effect observed with lysozyme is in stark contrast to myoglobin. The variability of the ATD increases almost uniformly, indicating that the protein is still unfolding and has not begun converging on a structure, and rather is continuing to unfold.

CIU fingerprint analysis

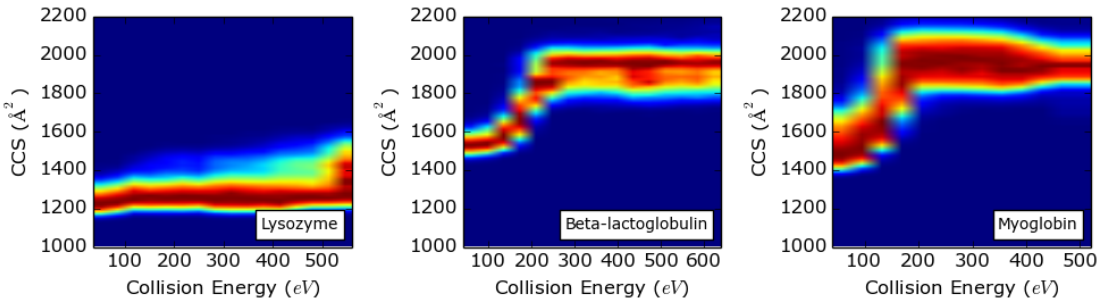

Figure 5.9: Collision induced unfolding (CIU) fingerprints of three model proteins; lysozyme (top), β-lactoglobulin (middle) and myoglobin (bottom).

CIU fingerprint analysis has also been carried out on the model proteins (Figure 5.9); in this case, the arrival time axis has been converted to CCS using Amphitrite. The general trends observed are similar to those found

using the summarising techniques; however, it is not possible to quantify them. This precludes automated analysis, as a user would need to manually inspect the data.

The Challenger program outputs tables of xy coordinates for the summary statistic results as well as figures, and a screenshot of the graphical user interface is shown in Figure 5.10. These tables can then be accessed by a script or opened in Microsoft Excel/LibreOffice Calc.

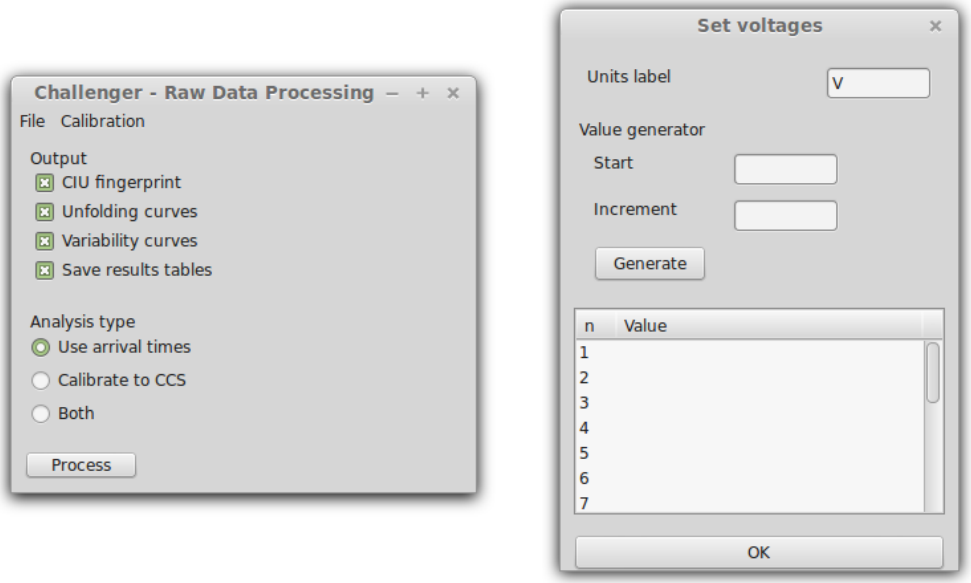

Figure 5.10: Graphical user interface for calculating and plotting summary statistics, as well as plotting CIU fingerprints. The program can also output the data as a table. An Amphitrite calibration can be added to convert arrival times to CCS. Main program is shown on left, with the dialog box for setting the voltages used in the unfolding experiment on the right.

These data can be directly accessed and so conditions can be decided for filtering samples. For example the selection criterion could be that the analyte has to have a smaller than 20 % increase in CCS at collision energies up to 300 eV. This criterion would select lysozyme and reject the other proteins in Figure 5.7C. This is similarly relevant to changes in conformational variability.

The summary statistics add a quantitative measure to changes in unfolding and variability which is not afforded by CIU analysis. This is beneficial in scenarios where there are small changes in the values as well as for automation.

The analysis is also objective and reproducible as no fitting is required. Additionally the analysis allows for the quantification of the differences in stability and conformational flexibility between analyte and the two complimenting techniques reduce the amount of data which has been lost.

5.3.3 Challenger algorithm development

Unfolding analysis of proteins often yields several transient conformational families [18]. Tracking the changes in abundance of these conformations during unfolding can be used for comparison between different proteins or proteins under different conditions. Thus far, peak heights have been used to represent the abundances of conformational families [14, 38]; here we describe a novel algorithm for the fitting and deconvolution of gas-phase unfolding arrival time distributions.

Ion mobility mass spectrometry instrumentation produces low resolution data, the result of this is that arrival time distributions conformations are rarely baseline separated. The poorly resolved data can cause problems when trying to fit Gaussian peaks using gradient descent optimisation techniques such as non-linear least squares [39, 40] as they are vulnerable to optimising to a local minima, which causes the algorithm to terminate [41]. For this reason a genetic algorithm approach was utilised which incorporates exploratory as well as exploitative optimisation [42].

A purpose built framework for deconvoluting unfolding data has been developed and is presented here. An important feature of proteins revealed in the ion mobility data is that, as a protein unfolds in the gas-phase, it shifts between discrete conformations and does not progressively unfold in a linear manner (this phenomenon can be seen in Suppl. Figure 5.5). This feature is used as a constraint in the fitting procedure; multiple ATDs are deconvoluted simultaneously but the peak centres of each conformation must stay constant across all ATDs. The method for creating the simulated data are shown in Equation 5.3, d is the number of data points per ATD, n the number of Gaussian distributions (conformations) being fit, S is the vector of simulated data points and μ, Γ and A are, respectively, the mean, full width half maximum

(FWHM) and amplitude of the Gaussian distribution.

$$S = \sum_{j=1}^{d} \sum_{i=1}^{n} A_i z^{-\frac{(x_{i,j}-\mu_i)^2}{2(\Gamma_i/2\sqrt{2ln2})^2}} \tag{5.3}$$

The fitness (F) of a chromosome is calculated as the additive inverse of the error. Two methods for calculating error are used here, the sum of absolute differences and sum of squares. The equation for calculating fitness based on these error values are shown in Equations 5.4 and 5.5 respectively, with a being the number of ATDs in the dataset.

The algorithm uses sum of squares error. This approach adds a greater penalty for large deviations, in comparison to sum of absolute differences, between experimental and simulated data allowing it to more aggressively optimise against larger problems in the solution. The result of this is that the algorithm reaches fits which are close to the original data within fewer generations. When used in figures (such as 5.13B) the axis is explicitly labelled to indicate that sum of squares error was used in the calculations.

$$F = -\sum_{k=1}^{a} \sum_{j=1}^{d} \left(\left| \sum_{i=1}^{n} A_{i,k} z^{-\frac{(x_{i,j,k}-\mu_i)^2}{2(\Gamma_{i,k}/2\sqrt{2\ln 2})^2}} - y_{j,k} \right| \right) \tag{5.4}$$

Sum of absolute differences is used when comparing to data derived from methods other than solutions generated by the algorithm, such as Figure 5.12 and is labelled 'Error'.

$$F = -\sum_{k=1}^{a} \sum_{j=1}^{d} \left(\sum_{i=1}^{n} \left(A_{i,k} z^{-\frac{(x_{i,j,k}-\mu_i)^2}{2(\Gamma_{i,k}/2\sqrt{2\ln 2})^2}} \right)^2 - y_{j,k}^2 \right) \tag{5.5}$$

Considerations of computation time

With the ability to automatically analyse the stabilising potential of putative ligands or drug molecules, the amount of processing time required to analyse the data becomes important. A gas-phase unfolding experiment can take approximately 30 minutes per sample, which means up to 48 samples could be run per day. The summary statistic analysis can be used to screen ligands to see if they stabilise/destabilise proteins as required, and the processing time is a matter of seconds. If a suitable ligand is found once per day, it would be problematic if the deeper analysis with the Challenger algorithm took weeks of processing time, this makes it important to reduce computation time as much as possible.

There is an existing Python library for running genetic algorithms called PyEvolve [29]; initially the algorithm was to be developed with this resource. It was quickly evident that the data processing would be too slow for use with our experimental data. When analysing the experimental data, it is beneficial to run the algorithm with different information regarding the peak centres (i.e. how many Gaussians to fit or the mean values for static mean analysis) to achieve an optimal solution.

In place of this, a purpose built genetic algorithm was developed in the C programming language [30]. The 32 fold speed increase is shown in Figure 5.11, 1,000 generations were computed using a simplified synthetic data set containing two ATDs comprising of 20 data points and 2 conformations when using dynamic means fitting. The calculation of error in our algorithm requires the majority of computation time. For each generation each individual (default 2,000), the error is determined by calculating the Gaussian distributions described by the individual and summing them, after this each simulated value for intensity is subtracted from the solution and absolute error is summed to give an error value.

The calculation lends itself to massively parallel solutions and GPGPU (General-Purpose computing on Graphics Processing Units) calculations so an implementation of the algorithm was created using Cuda [43]. For each individual in the population a GPU thread is spawned, which then spawns

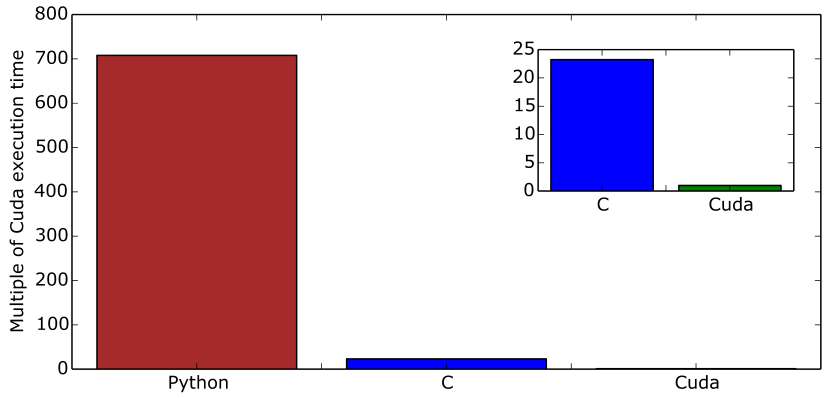

Figure 5.11: Comparison between execution times for the three implementations of the algorithm. The y axes are given as a proportion of the time taken to execute the Cuda version, for example the Python implementation was over 700 times slower. The inset is a magnified version of the main figure to emphasise the relationship between the C and Cuda implementations.

further threads for each data point in the arrival time axis. The calculation of each described Gaussian distribution, the summation to create a potential solution and the determination of absolute difference to the data can then be done in parallel potentially reducing computation time dramatically. The summing of the array result to produce the error is also parallelised, rather than taking the number of times it takes to calculate a sum multiplied by the number of data points (n) the calculation completes after $\log_2(n)$.

The development of the algorithm was done on an inexpensive consumer-level graphics card and the result of this parallelisation was a 23 fold reduction in computation time when timing 1,000 generations of evolution at a population size of 2,000 for experimental lysozyme data with 16 ATDs and 4 Gaussian distributions in comparison to the C implementation.

When collecting results based on the optimum parameters determined in Section 5.3.4 (population size 2,000, mutation rate 0.0001, uniform crossover, 200,000 generations) for lysozyme, the Cuda algorithm implementation takes approximately 1 hour to run, and so the equivalent time for the Python implementation would have been approximately 30 days.

Algorithm testing with synthetic data

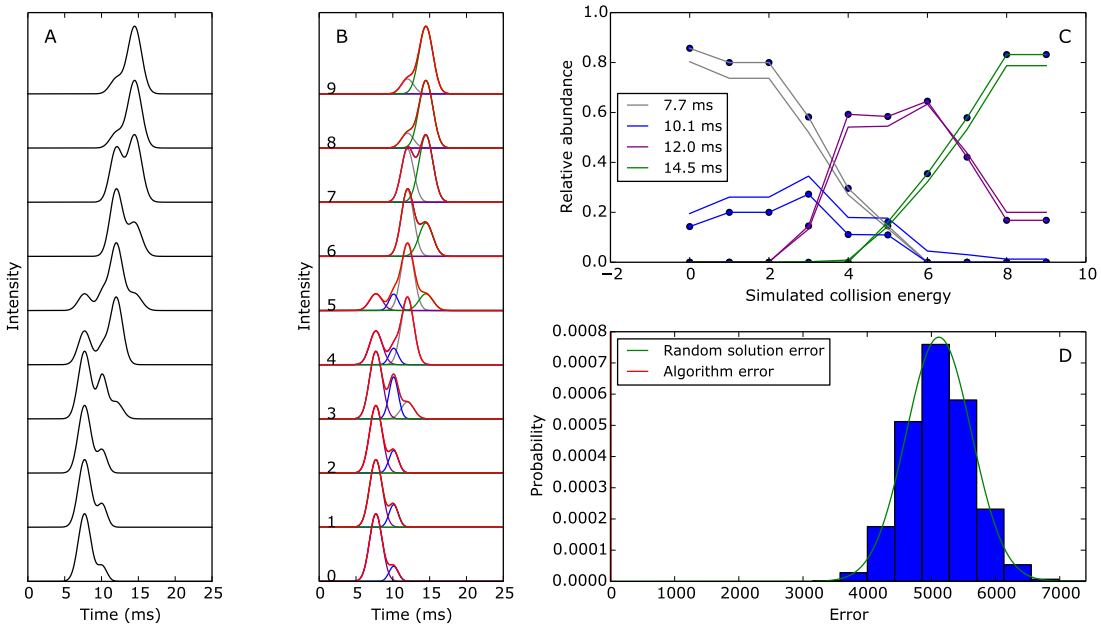

Figure 5.12: Deconvolution algorithm run on synthetic collision energy ramp data shown in (A). The deconvolution result (red) is shown in (B) with the original data as black traces and the individual conformations with other colours. The peak height (plain line) and peak area analysis (solid line with dot markers) for the data are shown in (C); In part (D) the error of 1 million randomly generated solutions is shown as a histogram and as a fitted Gaussian distribution in green, the error generated by the algorithm solution is the vertical red line.

In order to test the efficacy of the deconvolution methodology, a synthetic set of data representing an unfolding experiment was created (Figure 5.12A) and the deconvolution algorithm achieved very good agreement with the original data (Figure 5.12B).

In order to test the effectiveness of the algorithm against generating random solutions, 1 million random solutions were generated and are plotted on Figure 5.12C. The average error was calculated for the solution determined by the algorithm (3 replicates) and was plotted as a vertical red line.

The random solution error was tested for normality using the Anderson-Darling test [44], and by integrating above and below the algorithm error, it was determined that the probability of a random solution having the same error or less was 1.06×10^{-23}. Conversely the genetic algorithm only tested

in the order of 10^8 solutions before completing. The improvement using peak areas as opposed to peak heights for abundance analysis was also tested and the result is shown in Figure 5.12D.

5.3.4 Challenger algorithm optimisation

Crossover

Crossover is a method used to simulate the biological process of recombination. The genes of two parent chromosomes are mixed together to create the offspring. There are several different crossover implementations, and two common methods are compared here. In single point crossover a point along the chromosome is randomly chosen, the genes after that point are taken from one parent chromosome and the genes before from the other. The second method is uniform crossover where for each gene a random number between 0 and 1 is generated. Which parent the offspring's gene comes from is then determined by whether the number is above or below 0.5.

The single point crossover is most true to the biological process of recombination and only requires a single random number to be generated. In comparison for uniform crossover random numbers have to be generated, which can cause a significant increase to computation time and so it would be preferable to use single point crossover.

To test the two crossover methods, the lysozyme data set was run with a population size of 2,000, using three different mutation rates. The algorithm was run 32 times using different random number generator seeds to ensure that the different algorithm runs followed different evolutionary paths and the results are shown in Figure 5.13A.

Single point crossover outperformed uniform crossover at the highest mutation rate, however this was not maintained for lower mutation rates. The lowest average error achieved by single point crossover was 170 when a mutation rate of 0.001 was used. The error was higher than when using uniform crossover which achieved an error of 131 for the same mutation rate. Additionally the lowest average error achieved overall was uniform crossover at a

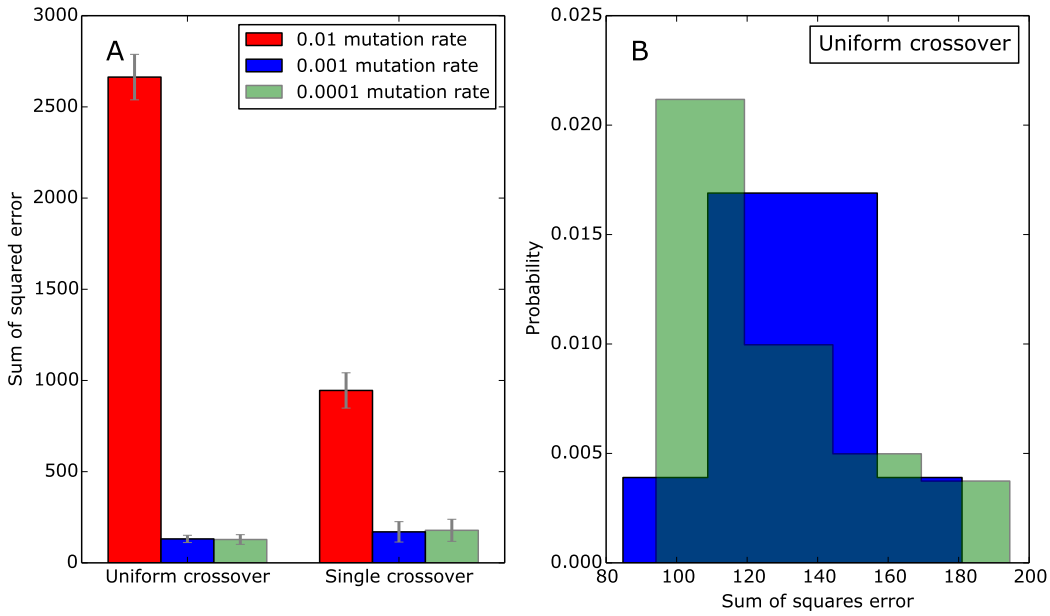

Figure 5.13: Optimising the mutation rate and crossover parameters, for experimental data. The algorithm was run 32 times for each combination of settings using the lysozyme dataset and a population size of 2,000, and the results are shown in (A). The error bars indicate the spread of 2 standard deviations. (B) Histogram comparing the error results for the two best performing configurations, uniform crossover with mutation rates of 0.001 and 0.0001.

mutation rate of 0.0001 (128), where single point crossover only achieved an error of 178.

Calculating the time difference between the methods showed that the single point crossover took 81 % of the time taken by the uniform method when all other factors were constant. This is in comparison to the percentage error difference where the single point method resulted in 33 % higher error when comparing the best result from each method. The additional time taken to run the uniform crossover is sufficiently beneficial to rule out the use of the single point crossover method. Users of the Challenger algorithm can pick which crossover implementation to use, but for the rest of the analysis in this thesis the uniform crossover method is used.

Mutation rate

Genetic algorithms utilise the biological process of genetic mutation to improve algorithm performance. Optimisation is required to determine an ap-

propriate mutation rate, which is the probability that a gene will be mutated each generation, and the implementation is that a new randomly generated number replaces the current value of the gene. High mutation rates mean that a larger array of values can be tested, which is especially important with small population sizes. Detrimentally, high mutation rates can disrupt the progression of the genetic algorithm by introducing too much randomness for the other simulated evolutionary processes to function.

Three mutation rates, 0.01 (1 % mutation chance), 0.001 and 0.0001, were compared and the results are shown in Figure 5.13A. The highest mutation rate performed very poorly, with much higher error values than the lower mutation rates. The total number of genes in each chromosome for this dataset was 132. This means that on average there would be 1.32 mutations per chromosome, which has been of detriment to the genetic algorithm's performance when searching for a solution.

The difference between mutation rates 0.001 and 0.0001 is closer, and though the average error is lower using lowest mutation rate, the standard deviation is higher. To further investigate this, a histogram was plotted (Figure 5.13B), and here we can see that though the variability is greater for the lowest mutation rate the distribution of the error values is better in comparison to the 0.001 mutation rate.

There is additionally a small time benefit for using lower mutation rates, as new values have to be generated less frequently. Upon testing, the 0.0001 mutation rate took 5 % less time to complete than the 0.001 mutation rate. For these reasons, the 0.0001 mutation rate has been used throughout the work presented in this thesis.

Population size

An important factor in the performance of a genetic algorithm is the population size, or number of chromosomes per generation. Larger population sizes improve the chance that the solution to a problem can be constructed from the initial population without having to rely on mutation to generate appropriate genes. This comes at the cost of the run time of the program as the amount of computation required scales linearly with the number of

chromosomes which are to be processed.

In order to select an appropriate population size to use for experimental data the lysozyme dataset was tested again. Uniform crossover was used with a 0.0001 mutation rate for a variety of population sizes. The algorithm was run for 200,000 generations and each result was replicated six times using different random number seeds, the results are shown in Figure 5.14.

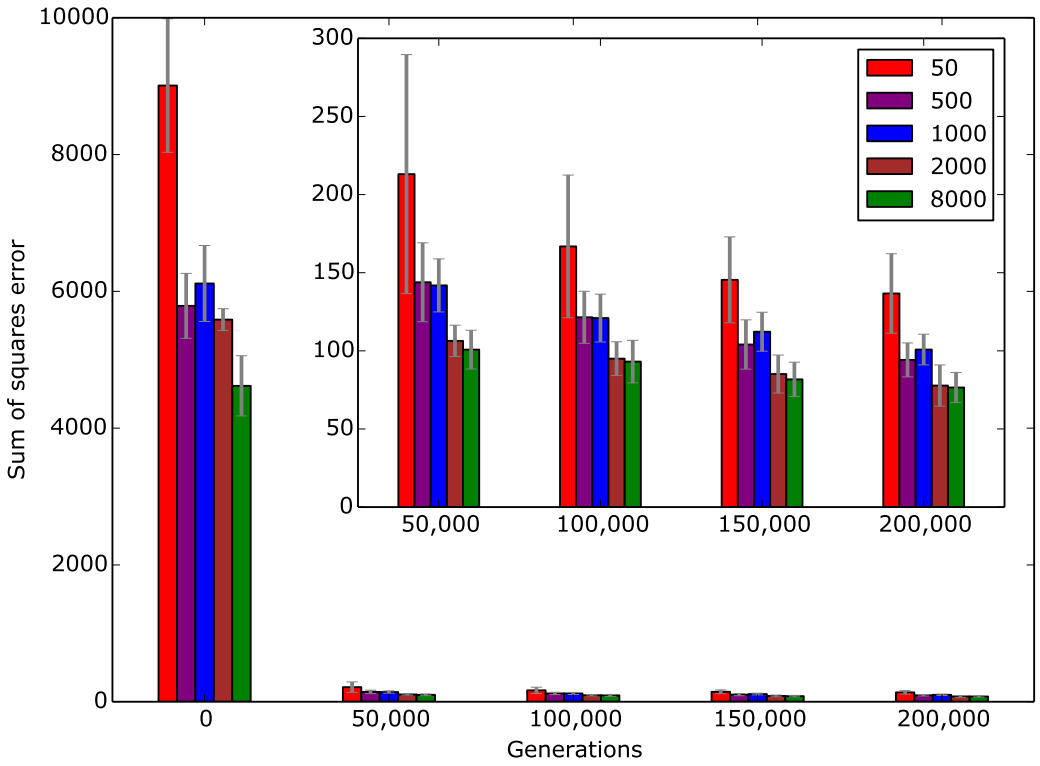

Figure 5.14: Optimising the population size for experimental lysozyme data. The algorithm was run for 200,000 generations and the error for each population size (as coloured in legend) was checked every 50,000 generations. The error bars show the spread of 2 standard deviations, and the inset is a magnification of the main figure.

The smaller population sizes, 50, 500 and 1,000, all perform substantially worse than 2,000 and 8,000. The difference between population sizes 2,000 and 8,000 is much smaller. The larger population does consistently perform better but it may not be worth the additional computation time.

Though the amount of computation required scales linearly to the population size with double the population size requiring twice the computation,

it may not be reflected in the increase in processing time as the Challenger algorithm has been massively parallelised using a GPU. Testing the time increase between population sizes of 2,000 and 8,000 showed that a population size of 2,000 takes 26.5 % of the time taken to process the larger population.

As a rough cost-benefit analysis, the error at 50,000 generations for 8,000 population size can be compared to the error at 200,000 generations for the 2,000 population size, as the computation time would be approximately equal. The average error for the smaller population was 76.5 against 100.8 for the population size of 8,000. This reason led to the use of a population size of 2,000 for the figures presented in this thesis, however the algorithm had to be run hundreds of times, so when trying to get the best fit possible, higher population sizes should still be considered.

Number of generations

The algorithm runs for a certain number of rounds of simulated evolution, with each round known as a generation. The error from one generation to the next never worsens, due to elitism, where the best chromosome from the last generation is retained.

Generation	Error	ΔError
0	5,587.1	n/a
50,000	106.4	-5,480.7
100,000	95.0	-11.4
150,000	85.1	-9.9
200,000	77.8	-7.4

Table 5.2: Table showing the average error and the change in error over the previous 50,000 generations for the population size of 2,000. data are taken from Figure 5.14.

The comparison of the effect of generations is shown in Figure 5.14. Generation 0 indicates the error for the fittest chromosome in the starting population, and the biggest reduction in error is shown in the first 50,000 generations. Table 5.2 shows the reduction in error from each additional set of 50,000 generations, and it shows that though the initial decrease in error is the largest, there is still consistent improvement until 200,000 generations. This indicates

that the algorithm could still achieve a lower error, and the result of this test is that when running the algorithm it is preferable to run it for as many generations as possible.

On the computer used for these calculations, it takes approximately 1 hour to run using the optimum settings shown here for 200,000 generations and 15 minutes for 50,000 generations.

5.3.5 Experimental data deconvolution

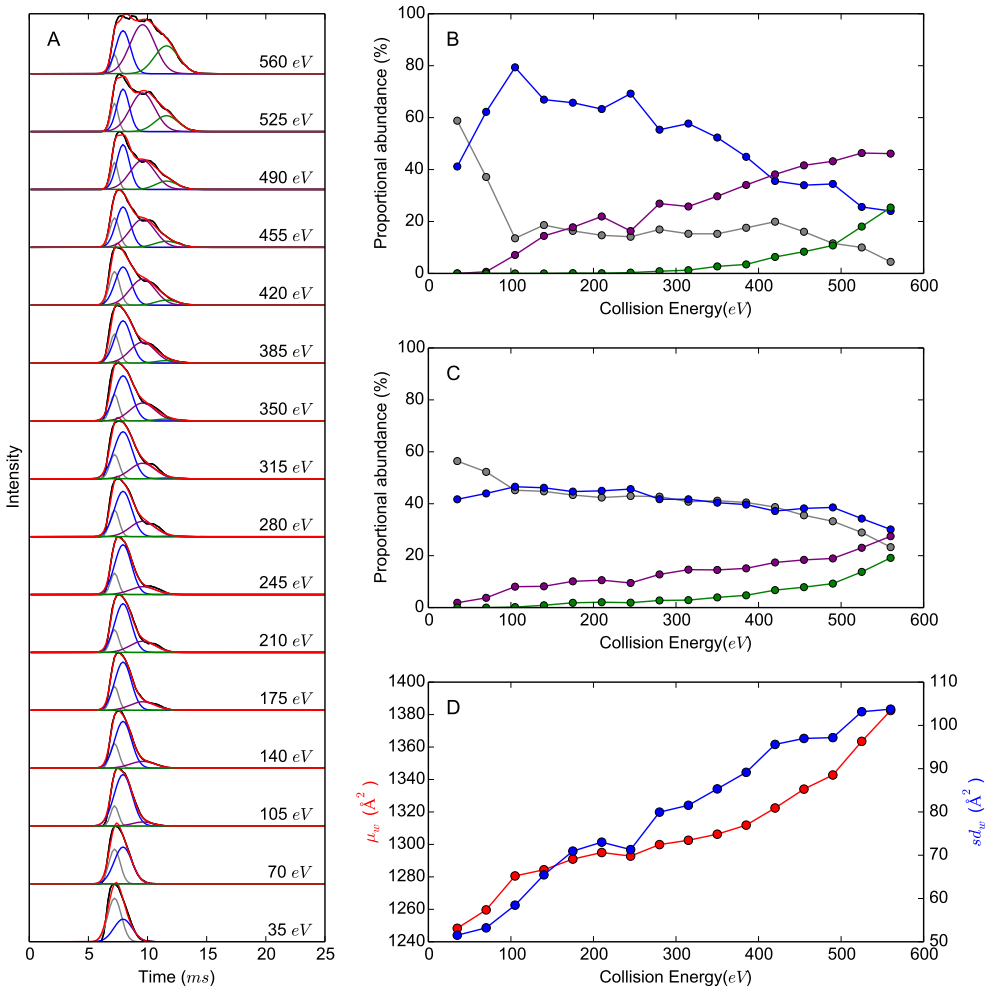

Figure 5.15: Deconvolution, abundance analysis and comparison to summary statistics of lysozyme gas-phase unfolding. (A) Deconvoluted ATDs, experimental data are shown in black and the simulation is red, individual conformations are coloured the same throughout the figure. (B) Conformational abundance analysis as area calculated from the deconvolution. (C) Conformational abundance determined using peak height analysis. (D) The results of the summary statistic analysis with the unfolding curve (weighted mean) in red and the variability curve (weighted standard deviation) in blue.

Using the optimised settings outlined in Section 5.3.4 and running the algorithm for 100,000 generations on the lysozyme dataset has yielded the data shown in Figure 5.15.

The deconvoluted arrival time distributions are shown in Figure 5.15A. All of the deconvoluted conformations summed together create simulated ATDs,

which are shown in red and show good agreement with the original experimental data, coloured black. The abundances of each conformation are displayed in Figure 5.15B, in terms of relative conformational abundance. That is the proportion of summed area of all conformations determined by the genetic algorithm for a particular ATD. Additionally the relative abundance as calculated from peak heights (Figure 5.15C) are shown with the results of summary statistics analysis shown in Figure 5.15D.

To assess the number of conformations which were used, the deconvoluted arrival time distributions should be inspected. An indication that there have not been enough conformations fitted is that there are intensity regions in the arrival time distribution which are not present in the simulated data. An example of this is the extended conformation, which appears as a peak shoulder, seen in Figure 5.16A. The opposing scenario is when several simulated conformations are fit to a single experimental feature and would be known as overfitting. An example of overfitting is shown in Figure 5.16B, where there are two or three conformations present in the data, but eight conformations have been fit.

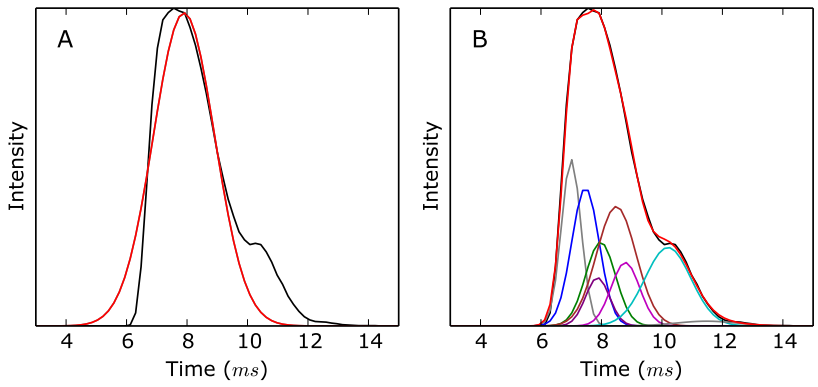

Figure 5.16: Examples of an experimental ATD (black) with poor fits. (A) A single conformation is fit resulting in a large difference in the simulated (red) and experimental data. (B) The data has been fit with 8 conformations, though the simulated and experimental data are very similar, there are many individual conformations (other colours) fitting to each single feature in the experimental data, which is known as overfitting and is unlikely to be a true representation of the conformational families that combine to make an experimental ATD.

The deconvolution shown in Figure 5.15 simulated four conformations. By

examining the Figure 5.15A, it can be seen that there are no unidentified features in the ATDs, and there are not multiple very close together or low intensity conformations. Thus we can conclude that this is a good fit.

The conformational abundance analysis from the lysozyme deconvolution, using deconvoluted peak areas and peak top intensity are shown in Figure 5.15B and C respectively. The data show that the most closed conformation (grey) is overrepresented consistently after the lowest voltage, and is due to the conformation overlapping with a more open conformation (blue). At the highest collision energy, the blue conformation is represented the most abundant, whereas the deconvolution shows that it is the second scarcest conformation.

The unfolding and variability curves are included in the figure (Figure 5.15D) to compare with the results for conformational abundance via deconvolution (Figure 5.15B). In Section 5.3.2, it was proposed that the continually increasing level of unfolding and conformational variability could be due to the protein occupying more conformations as the collision energy increased, and that as the variability was still high, the protein would not have converged on a particular structure. The conformational abundance analysis shows these statements to be true, as the collision energy increases, additional conformations are detected (purple and green). The protein has not converged on a single conformation by the end of the collision energy ramp, with four conformations being present, three of which make up over 20 % each of the total abundance, and the most abundant conformation makes up less than 50 % of the total area of the ATD.

5.3.6 Algorithm limitation

The determination of peak centres by the algorithm is not always perfect. Difficulties arise when many ATDs are very similar, the result of this is that there is a bigger advantage to solving the fit for that region as opposed to areas of intensity which only occur in a few ATDs. An example of this is shown using the β-lactoglobulin dataset in Figure 5.17.

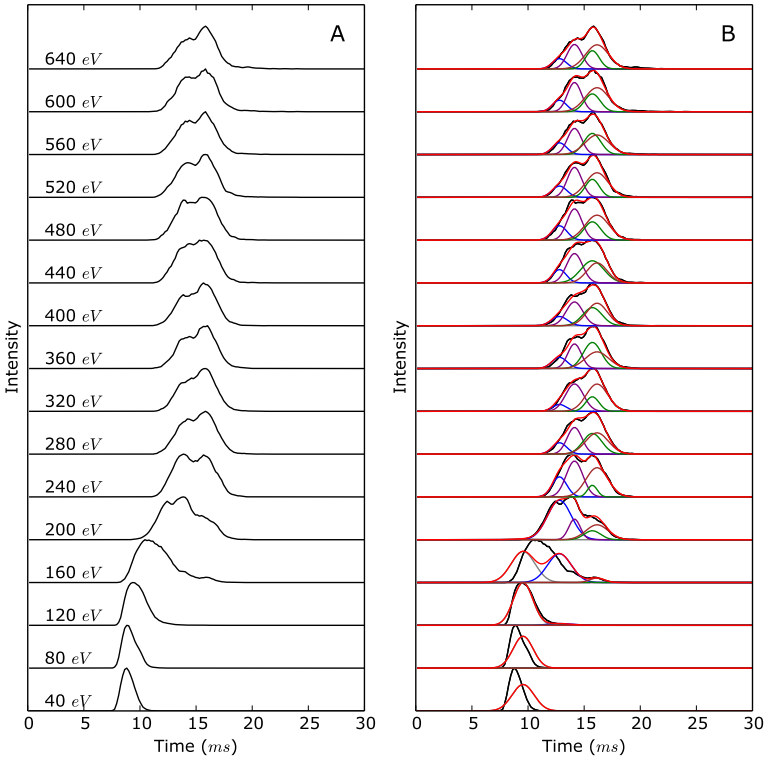

Figure 5.17: Demonstration of a limitation of the deconvolution algorithm when analysing data with many similar ATDs. (A) The β-lactoglobulin unfolding dataset as ATDs. (B) The deconvolution result, with experimental data shown in black, simulated data in red, and individual conformations using different colours.

With this dataset, 10 of the 16 ATDs are very similar (280-640 eV), and the algorithm has fitted these ATDs very well. The four lowest collision energy ATDs look to require three conformations in order for them to be fit correctly, and the algorithm has only assigned one (grey).

The summary statistic results in Section 5.3.2 show that there is little change to the degree of unfolding or conformational variability after 300 eV. In conjunction with the incidence of very visually similar ATDs after that point, we can remove some of those ATDs from the analysis without losing information.

The ATDs from 400 eV and above were removed from the dataset, and the Challenger algorithm was run again. The results are shown in Figure 5.18, and the fitting to the ATDs is much better. There is no overfitting in any areas, and all key features of the dataset are described appropriately with

simulated conformations.

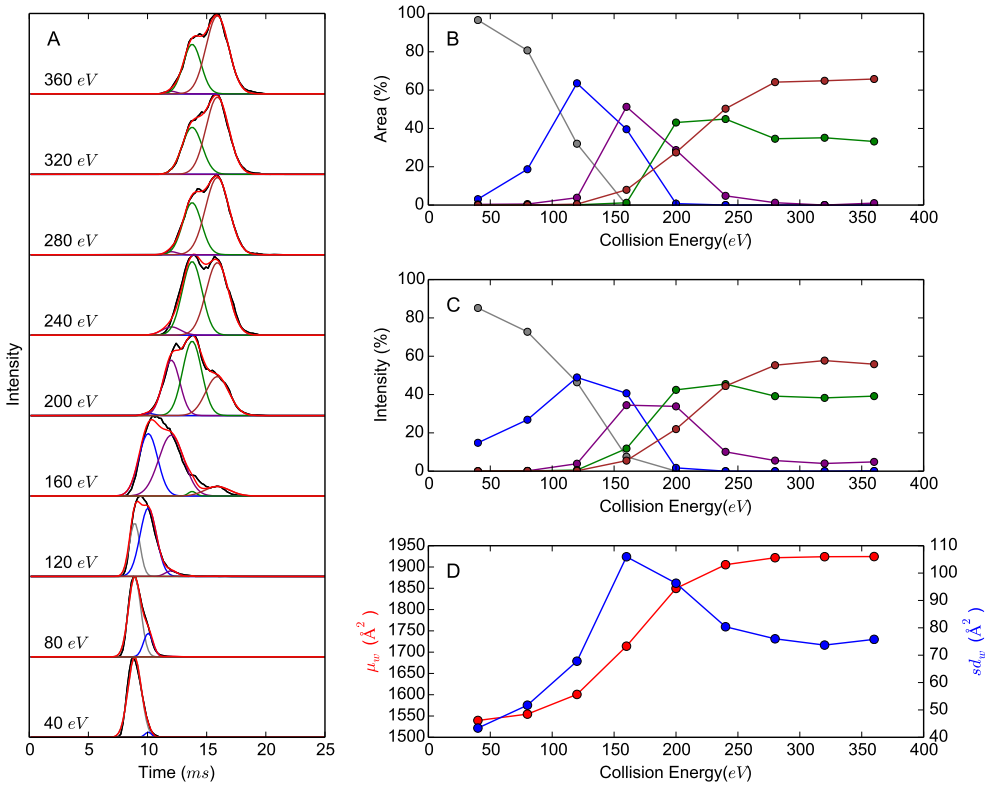

Figure 5.18: Deconvolution of a reduced β-lactoglobulin dataset. (A) Deconvoluted arrival time distributions coloured as in Figure 5.17. Abundance analysis using deconvoluted areas of each conformation (B) and the peak top intensity (C). (D) Summary statistic analysis of the dataset with unfolding curve shown in red and variability curve in blue.

The comparison between peak areas (Figure 5.18B) and heights (Figure 5.18C), once again show that overlapping peaks cause distortion in the reported abundance for the peak height method. This is evident in the comparison between the two most closed conformations (grey and blue) at lower energies. The variability curve of β-lactoglobulin (Figure 5.18D), shows rises to a peak at 100-200 eV, and then falls. This phenomenon would be predicted to be caused by an unfolding event, where multiple species are present, before converging towards fewer conformations. This can be confirmed by examining Figure 5.18B, at low collision energies there is a single predominant conformation, at 100-200 eV there are four conformations present. Finally, the protein converges to two conformations, with the three most closed conformations having abundances near to 0 %.

5.3.7 Current software development state

The summary statistic algorithms, originally accessed as a Python library, has now had a graphical user interface (GUI) developed for it. This allows users to quickly and easily create CIU fingerprints, a table of results and the unfolding and variability curves introduced in this chapter and can be seen in Figure 5.19. The table of results gives the output of the summary statistic calculations so that users can create plots with their favourite plotting program. The data can be calibrated to CCS using an Amphitrite calibration file and data can be provided in comma separated value (CSV) format.

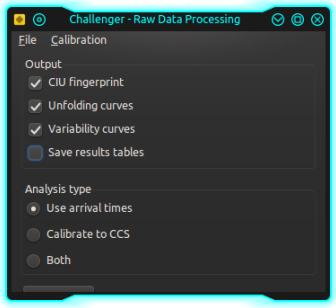

Figure 5.19: Graphical user interface for calculating summary statistics and CIU fingerprints from unfolding data.

For the Challenger algorithm, users can create a simple Python script in order to run the analysis. There are several input parameters and so a GUI has not yet been created. An example script is shown below.

```
import challengerClasses.RunChallenger as RunChallenger
# ================
# Setting input variables

popSize = # Population size
nGens = # Number of generations to run
nGauss = # The number of conformations to be fit
nAtds = # Number of ATDs in the input data file
repBak = # How often, in generations, the algorithm should report back progress
crossover = # 0 for uniform crossover, 1 for one point cross over
mutationRate = # Mutation rate e.g. 0.01 for 1 % of alleles mutated each generation

dataFn = # Filename/path of input data
outputDir = # Path to directory to store results
seed = # Random number generation seed. False to not use a seed.

# ================
# Setup, compile and run
ob = RunChallenger.ChallengerSetup(popSize,nGens,nGauss,nAtds,repBak,crossover,mutationRate)
ob.compile()
ob.run(dataFn,outputDir,seed)
```

5.4 Conclusion

This chapter has introduced a method of summarising an ATD to two values, weighted mean and weighted standard deviation. This approach has been shown to be more representative of the original data than using the peak apex method, by including information about secondary conformations and the amount of total conformational variability. This is achieved whilst still maintaining the quantitative nature of the data. This contrasts to the recently introduced CIU fingerprint technique for analysing unfolding data, which is a qualitative visual data representation technique, potentially leading to subjectivity in analysis. CIU fingerprints are still a complimentary technique to summary statistic analysis, so the software presented is able to automatically generate figures for CIU fingerprints, unfolding curves and variability curves.

A method of automating the data acquisition for gas-phase unfolding experiments has also been introduced. This allows the instrument to be used continuously, acquiring unfolding data for many analyte mixtures. This makes the tool amenable to high-throughput drug discovery experiments. The summary statistic analysis is well suited to this environment due to its quantitative nature. Target values for degree of unfolding (ΔCCS) and variability can be set for different collision energies. During the high-throughput screening process, each new ligand could be analysed automatically. If a ligand meets the target values set, then it could be selected to be examined in greater detail using the Challenger algorithm.

The Challenger algorithm presented here is the first program developed for the deconvolution of IM-MS arrival time distributions. The results have shown that it allows for a more accurate determination of the abundances of individual conformations than using peak intensity. The algorithm also computes the centre of each conformation, which is likely more accurate and reproducible than operators selecting peak apexes by eye.

The results of the deconvolution can be used in a wider variety of data analysis applications than have been covered here. One such example would be the deconvolution of CIU fingerprints as shown in Appendix 5.22. This

gives a visual representation of the conformational variability (as width) of each conformation as well as intensity during the course of the unfolding experiment. It is preferable to the standard CIU data representation technique as each conformation can be seen clearly and overlapping conformations do not exhibit increased colour intensity.

It has been shown that the limitation of the Challenger algorithm covered in Section 5.3.6 can be overcome by removing large numbers of highly similar ATDs. A potential future solution for this problem would be to break the deconvolution process into two parts. The set of ATDs could be grouped using a similarity method, and a single ATD would be taken from each group. These ATDs would then be analysed using the standard Challenger algorithm, giving accurate determination of the centres of conformations present. The analysis would be fast as there would be less data to analyse. With the peak centres determined, a second round of deconvolution using the whole set of ATDs would be performed where the peak centres are not fit. This would greatly reduce the potential combinations of genes, leading to shorter computation times in comparison to the full Challenger algorithm. As a proof of principle the full β-lactoglobulin dataset was deconvoluted using static peak centres and the results can be seen in Appendix 5.21. The approach fixed the issue with all the simulated and experimental data being very similar and the second round of deconvolution only took 15 minutes to complete.

References

[1] Scarff, C. A., Thalassinos, K., Hilton, G. R., and Scrivens, J. H. (2008). "Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements". *Rapid Communications in Mass Spectrometry* 22.20, pp. 3297–3304.

[2] Ruotolo, B. T., Giles, K., Campuzano, I., Sandercock, A. M., Bateman, R. H., and Robinson, C. V. (2005). "Evidence for macromolecular protein rings in the absence of bulk water". *Science* 310.5754, pp. 1658–1661.

[3] Leary, J. A., Schenauer, M. R., Stefanescu, R., Andaya, A., Ruotolo, B. T., Robinson, C. V., Thalassinos, K., Scrivens, J. H., Sokabe, M., and Hershey, J. W. (2009). "Methodology for measuring conformation of solvent-disrupted protein subunits using T-WAVE ion mobility MS: an investigation into eukaryotic initiation factors". *Journal of the American Society for Mass Spectrometry* 20.9, pp. 1699–1706.

[4] Shelimov, K. B., Clemmer, D. E., Hudgins, R. R., and Jarrold, M. F. (1997). "Protein structure in vacuo: Gas-phase conformations of BPTI and cytochrome *c*". *Journal of the American Chemical Society* 119.9, pp. 2240–2248.

[5] Clemmer, D. E. and Jarrold, M. F. (1997). "Ion mobility measurements and their applications to clusters and biomolecules". *Journal of Mass Spectrometry* 32.6, pp. 577–592.

[6] Bernstein, S. L., Liu, D., Wyttenbach, T., Bowers, M. T., Lee, J. C., Gray, H. B., and Winkler, J. R. (2004). "α-synuclein: Stable compact and extended monomeric structures and pH dependence of dimer formation". *Journal of the American Society for Mass Spectrometry* 15.10, pp. 1435–1443.

[7] Valentine, S. J., Anderson, J. G., Ellington, A. D., and Clemmer, D. E. (1997). "Disulfide-intact and -reduced lysozyme in the gas phase: confor-

mations and pathways of folding and unfolding". *The Journal of Physical Chemistry B* 101.19, pp. 3891–3900.

[8] Teplow, D. B., Lazo, N. D., Bitan, G., Bernstein, S., Wyttenbach, T., Bowers, M. T., Baumketner, A., Shea, J.-E., Urbanc, B., Cruz, L., Borreguero, J., and Stanley, H. E. (2006). "Elucidating amyloid β-protein folding and assembly: a multidisciplinary approach". *Accounts of Chemical Research* 39.9, pp. 635–645.

[9] Bernstein, S. L., Wyttenbach, T., Baumketner, A., Shea, J.-E., Bitan, G., Teplow, D. B., and Bowers, M. T. (2005). "Amyloid β-Protein: Monomer Structure and Early Aggregation States of Aβ42 and Its Pro19 Alloform". *Journal of the American Chemical Society* 127.7, pp. 2075–2084.

[10] McLafferty, F. W. and Bryce, T. A. (1967). "Metastable-ion characteristics: characterization of isomeric molecules". *Chemical Communications* 23, pp. 1215–1217.

[11] Jennings, K. R. (1968). "Collision-induced decompositions of aromatic molecular ions". *International Journal of Mass Spectrometry and Ion Physics* 1.3, pp. 227–235.

[12] Hopper, J. T. and Oldham, N. J. (2009). "Collision induced unfolding of protein ions in the gas phase studied by ion mobility-mass spectrometry: the effect of ligand binding on conformational stability". *Journal of the American Society for Mass Spectrometry* 20.10, pp. 1851–1858.

[13] Fändrich, M., Forge, V., Buder, K., Kittler, M., Dobson, C. M., and Diekmann, S. (2003). "Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments." *Proceedings of the National Academy of Sciences* 100.26, pp. 15463–8.

[14] Hyung, S.-J., Robinson, C. V., and Ruotolo, B. T. (2009). "Gas-phase unfolding and disassembly reveals stability differences in ligand-bound multiprotein complexes". *Chemistry & Biology* 16.4, pp. 382–390.

[15] Hoaglund-Hyzer, C. S., Counterman, A. E., and Clemmer, D. E. (1999). "Anhydrous protein ions". *Chemical Reviews* 99.10, pp. 3037–3080.

[16] Pringle, S. D., Giles, K., Wildgoose, J. L., Williams, J. P., Slade, S. E., Thalassinos, K., Bateman, R. H., Bowers, M. T., and Scrivens, J. H. (2007). "An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument". *International Journal of Mass Spectrometry* 261.1, pp. 1–12.

[17] Rabuck, J. N., Hyung, S.-J., Ko, K. S., Fox, C. C., Soellner, M. B., and Ruotolo, B. T. (2013). "Activation state-selective kinase inhibitor assay based on ion mobility-mass spectrometry". *Analytical Chemistry* 85.15, pp. 6995–7002.

[18] Ruotolo, B., Hyung, S.-J., Robinson, P., Giles, K., Bateman, R., and Robinson, C. (2007). "Ion mobility–mass spectrometry reveals long-lived, unfolded intermediates in the dissociation of protein complexes". *Angewandte Chemie International Edition* 46.42, pp. 8001–8004.

[19] Wojdyr, M. (2010). "Fityk: a general-purpose peak fitting program". *Journal of Applied Crystallography* 43.5, pp. 1126–1128.

[20] Wojnowska, M., Yan, J., Sivalingam, G. N., Cryar, A., Gor, J., Thalassinos, K., and Djordjevic, S. (2013). "Autophosphorylation Activity of a Soluble Hexameric Histidine Kinase Correlates with the Shift in Protein Conformational Equilibrium". *Chemistry & Biology* 20.11, pp. 1411–1420.

[21] Jenner, M., Ellis, J., Huang, W.-C., LloydRaven, E., Roberts, G. C. K., and Oldham, N. J. (2011). "Detection of a Protein Conformational Equilibrium by Electrospray Ionisation-Ion Mobility-Mass Spectrometry". *Angewandte Chemie International Edition* 50.36, pp. 8291–8294.

[22] Giles, K., Pringle, S. D., Worthington, K. R., Little, D., Wildgoose, J. L., and Bateman, R. H. (2004). "Applications of a travelling wave-based radio-frequency-only stacked ring ion guide". *Rapid Communications in Mass Spectrometry* 18.20, pp. 2401–2414.

[23] Sivalingam, G. N., Yan, J., Sahota, H., and Thalassinos, K. (2013). "Amphitrite: A program for processing travelling wave ion mobility

mass spectrometry data". *International Journal of Mass Spectrometry* 345–347, pp. 54–62.

[24] Hernández, H. and Robinson, C. V. (2007). "Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry." *Nature protocols* 2.3, pp. 715–726.

[25] Bush, M. F., Hall, Z., Giles, K., Hoyes, J., Robinson, C. V., and Ruotolo, B. T. (2010). "Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology." *Analytical chemistry* 82.22, pp. 9557–65.

[26] Thalassinos, K., Grabenauer, M., Slade, S. E., Hilton, G. R., Bowers, M. T., and Scrivens, J. H. (2009). "Characterization of phosphorylated peptides using traveling wave-based and drift cell ion mobility mass spectrometry." *Analytical chemistry* 81.1, pp. 248–54.

[27] Mitchell, M. (1998). *An Introduction to Genetic Algorithms (Complex Adaptive Systems)*. The MIT Press, p. 221.

[28] Van Rossum, G. and Et Al. (2010). "The Python programming language". *Python Software Foundation*.

[29] Perone, C. S. (2009). "Pyevolve: a Python Open-Source Framework for Genetic Algorithms". *SIGEVOlution* 4.1, pp. 12–20.

[30] Brian W Kernighan, D. M. R. (1988). *The C programming language*. Prentice Hall.

[31] Sanders, J. and Kandrot, E. (2010). *CUDA by Example: An Introduction to General-Purpose GPU Programming*. Addison-Wesley Professional.

[32] Oliphant, T. (2007a). "Python for Scientific Computing". *Computing in Science & Engineering* 9.3.

[33] Oliphant, T. E. (2007b). "SciPy: Open source scientific tools for Python". *Computing in Science and Engineering* 9.3, pp. 10–20.

[34] Hunter, J. (2007). "Matplotlib: A 2D graphics environment". *Computing in Science & Engineering* 9.3.

[35] Zhong, Y., Hyung, S.-J., and Ruotolo, B. T. (2011). "Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions." *The Analyst* 136.17, pp. 3534–41.

[36] Wetzel, R., Perry, L. J., Baase, W. A., and Becktel, W. J. (1988). "Disulfide bonds and thermal stability in T4 lysozyme." *Proceedings of the National Academy of Sciences* 85.2, pp. 401–405.

[37] Burova, T. V., Choiset, Y., Tran, V., and Haertlé, T. (1998). "Role of free Cys121 in stabilization of bovine beta-lactoglobulin B." *Protein Engineering* 11.11, pp. 1065–1073.

[38] Freeke, J., Bush, M. F., Robinson, C. V., and Ruotolo, B. T. (2012). "Gas-phase protein assemblies: Unfolding landscapes and preserving native-like structures using noncovalent adducts". *Chemical Physics Letters* 524, pp. 1–9.

[39] Levenberg, K. (1944). "A method for the solution of certain non-linear problems in least squares". *The Quarterly of Applied Mathematics* 2, pp. 196–168.

[40] Marquardt, D. W. (1963). "An algorithm for least-squares estimation of nonlinear parameters". *Journal of the Society for Industrial and Applied Mathematics* 11.2, pp. 431–441.

[41] Liu, G., Han, X., and Lam, K. (2002). "A combined genetic algorithm and nonlinear least squares method for material characterization using elastic waves". *Computer Methods in Applied Mechanics and Engineering* 191.17, pp. 1909–1921.

[42] Alba, E. and Dorronsoro, B. (2005). "The exploration/exploitation tradeoff in dynamic cellular genetic algorithms". *IEEE Transactions on Evolutionary Computation* 9.2, pp. 126–142.

[43] Pospichal, P., Jaros, J., and Schwarz, J. (2010). “Parallel Genetic Algorithm on the CUDA Architecture”. *Applications of Evolutionary Computation*. 6024. Springer Publishing.

[44] Anderson, T. W. and Darling, D. A. (1952). “Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes”. *The Annals of Mathematical Statistics*, pp. 193–212.

5.5 Appendix

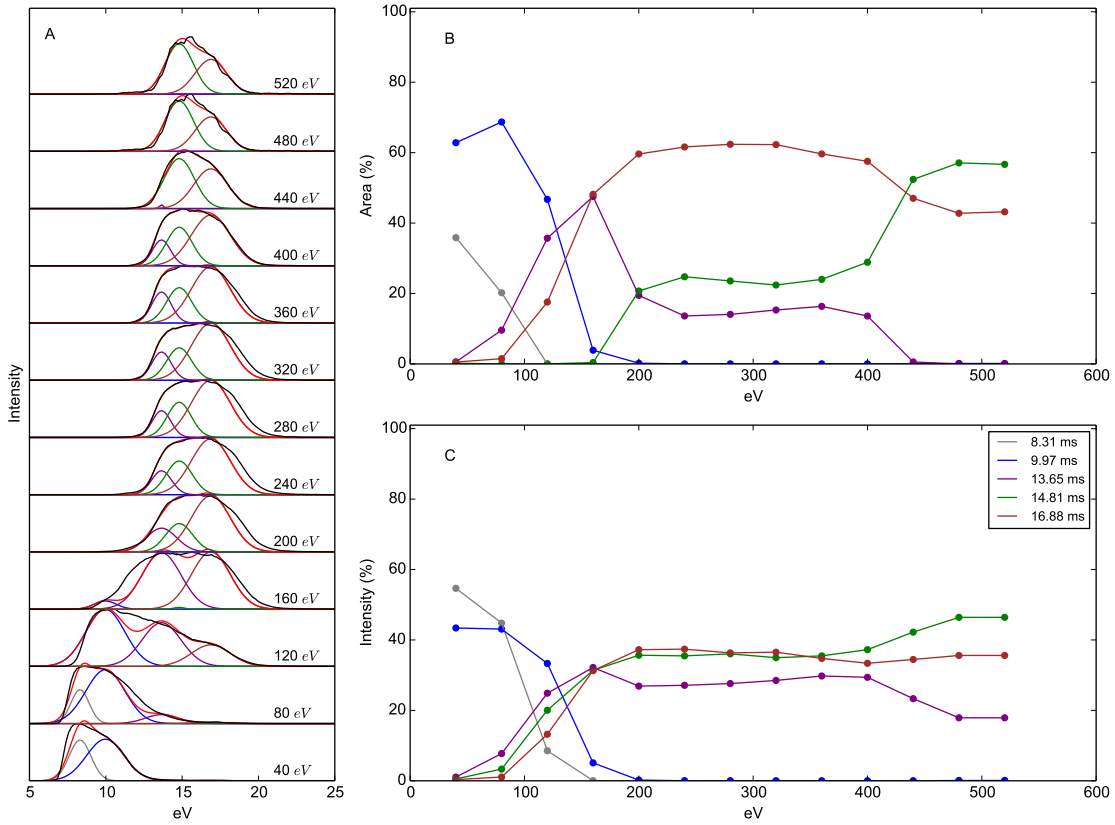

Figure 5.20: Challenger algorithm deconvolution of myoglobin.

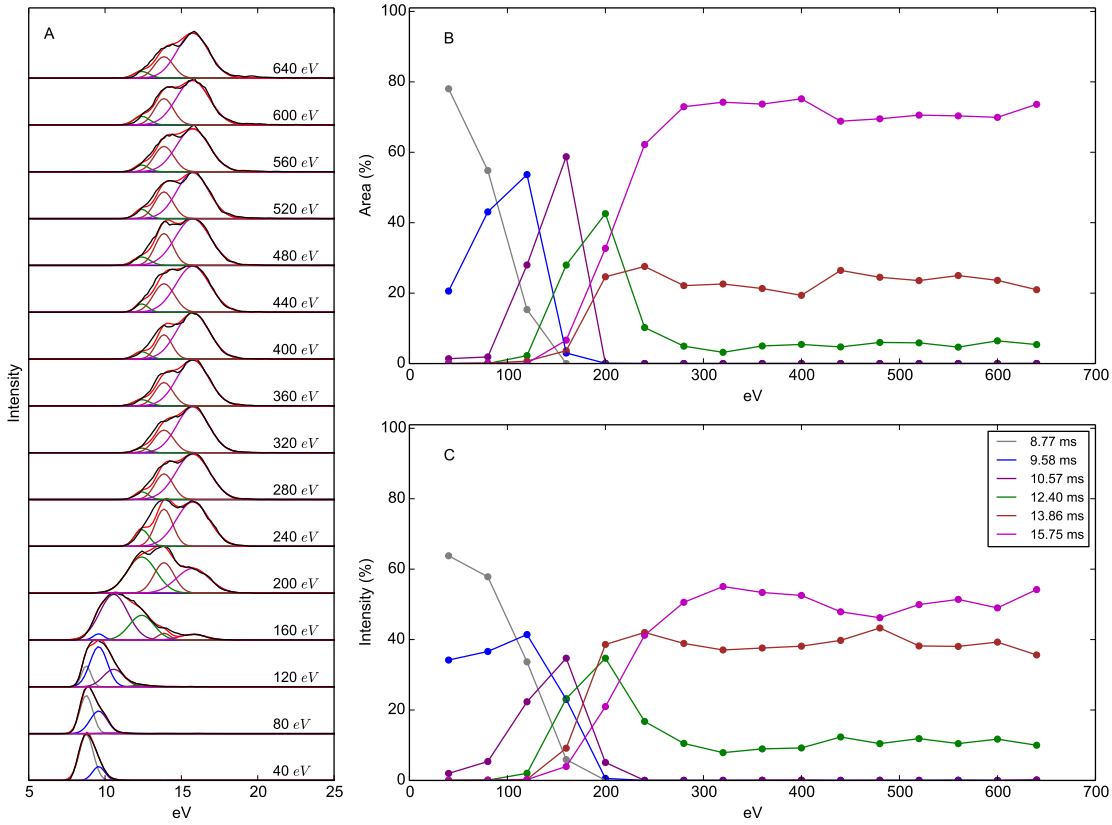

Figure 5.21: Deconvolution of β-lactoglobulin data when not fitting for conformational centres.

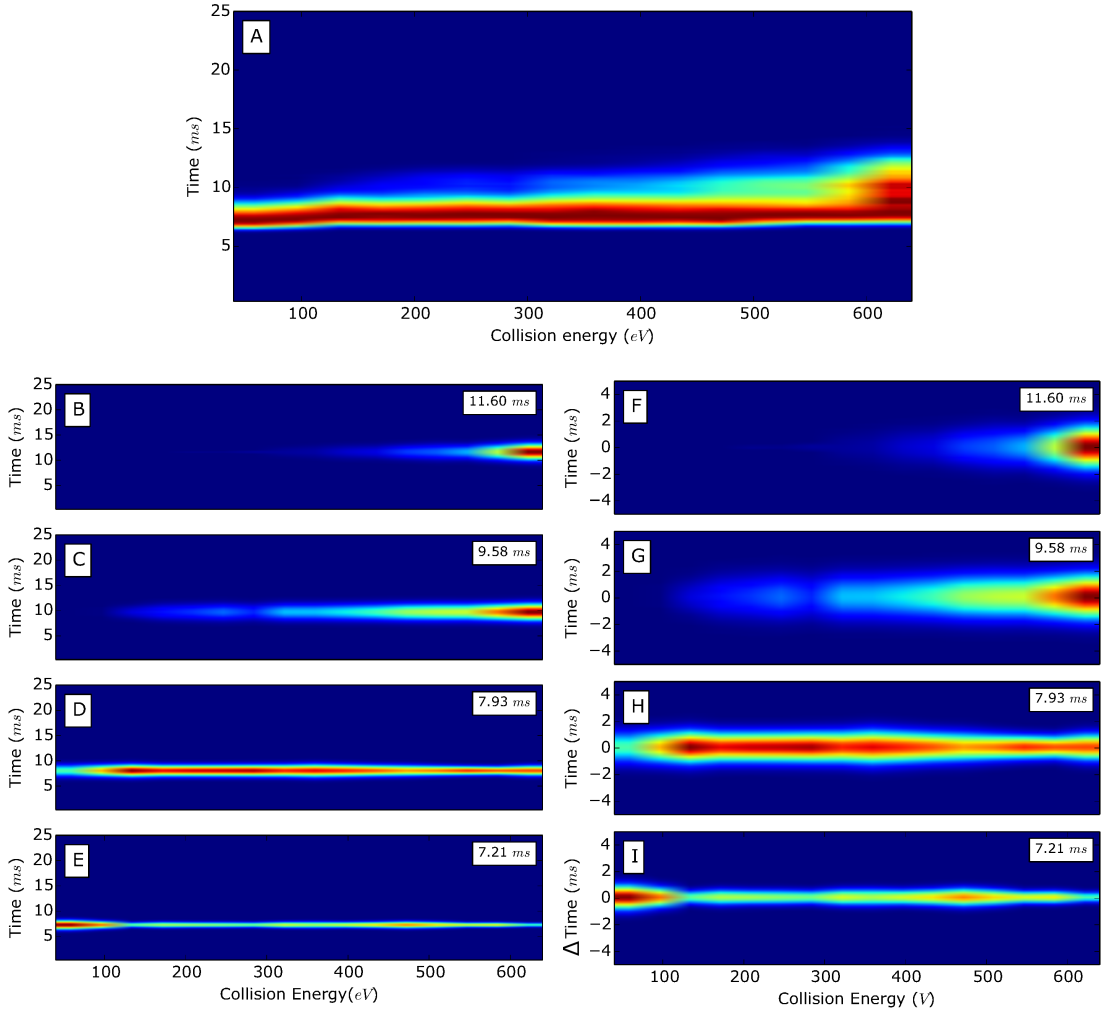

Figure 5.22: Using results of the Challenger algorithm to deconvolute lysozyme CIU fingerprints. (A) The CIU fingerprint. (B-E) Deconvoluted CIU fingerprints, with plot for each conformation (the conformation centre is labelled). (F-I) The same data as (B-E), but the y axis is the change in arrival time from the centre of the conformation, which is useful for observing the conformational variability.

Chapter 6

α_1-antitrypsin

This thesis has introduced several new method of the data analysis and representation of ion mobility mass spectrometry data. To demonstrate the efficacy of these methods, they have been applied to new research on the α_1-antitrypsin system here.

6.1 Introduction

α_1-antitrypsin is a member of the serpin (serine protease inhibitor) superfamily. The superfamily is homologous with similar tertiary structures; the proteins are composed of three β sheets, nine alpha helices and a reactive centre loop (RCL) as shown in Figure 6.1 [2]. The proteins function by binding a protease to the reactive centre loop where one end of the RCL is cleaved from the protein by the protease between the $P_1 - P_1'$ residues and the P_1 residue forms a covalent bond with the substrate enzyme. The reactive loop is then free to move and inserts into the central position in β sheet A between the two parallel strands 3 and 5 (s3A and s5A) as a strand in the anti-parallel direction (s4A) completing the anti-parallel β sheet [3]. This mechanism concomitantly pulls the substrate to the opposite end of the serpin sheet which disrupts the catalytic site and inhibits the activity of the protease. The apo protein is metastable with thermal stability up to 50-60 °C, conversely the

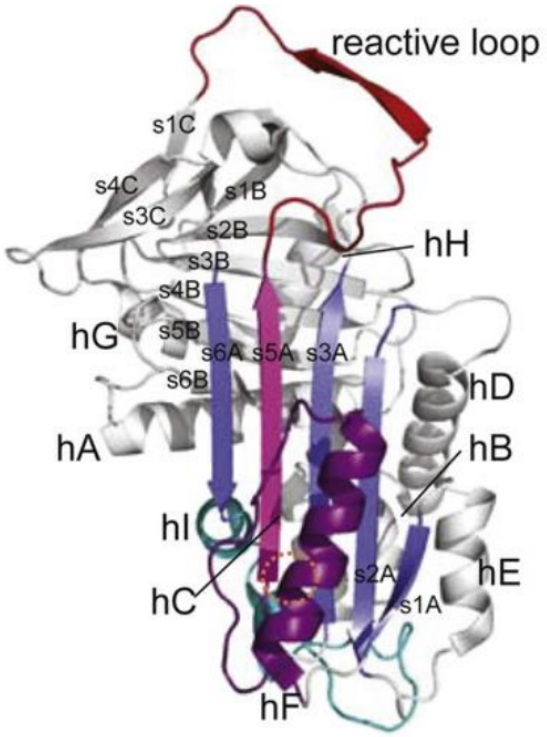

Figure 6.1: X-ray crystallography structure of α_1-antitrypsin, with secondary structure elements labelled. Helices are labelled hx where x indicates which helix. Strands are labelled with the format snx with x representing the β-sheet the strand belongs to and n being the strand number. Figure adapted from [1].

inhibitor-substrate complex is stable to over 100 °C [2]. These proteins are of specific research interest due to the ability to form highly stable aggregates due to point mutations. In this field aggregation is referred to as polymerisation and the disease state is caused by the homopolymerisation of serpins and the diseases are collectively referred to as the serpinopathies [4].

6.1.1 Pathology of severe α_1-antitrypsin deficiency

α_1-antitrypsin is the most abundant protease inhibitor found in plasma under normal conditions (1-2 mg/ml) [5, 6]. It was first discovered in 1955 [7] and was linked to emphysema and liver disease in 1963 [8] and 1969 [9] respectively. These features are now known to be caused by polymerisation of α_1-antitrypsin.

α_1-antitrypsin is predominantly synthesised in the endoplasmic reticula

of hepatocytes. In the case of severe α_1-antitrypsin deficiency, the protein forms linear polymer chains which then form Schiff positive inclusions bodies, which accumulate within the hepatocyte compromising function and inducing cell death [2]. The increased apoptotic rate of hepatocytes can then lead to cirrhosis of the liver [9] and neonatal hepatitis [10].

The primary target of α_1-antitrypsin is neutrophil elastase, an acute phase serpin protease which can kill gram-negative bacteria [11]. However its broad specificity means that it can also attack host tissue [11]. Without the inhibitory protection of α_1-antitrypsin due to retention in hepatocytes the uncontrolled proteolytic activity can cause damage to lung tissue leading to early onset panlobular emphysema [12].

6.1.2 Naming of α_1-antitrypsin variants

It was discovered early on that there were numerous α_1-antitrypsin variants, and so the Pi (protease inhibitor) test was created to determine what the effect of different variants would be [13]. The variants are given letters dependent on their mobility in isoelectric focusing gel electrophoresis; F for fast mobility, M for medium, S for slow and Z for very slow [13–15]. F bands are rare and are caused by a more closed conformation of the protein, and patients carrying such a form of α_1-antitrypsin can also develop lung disease [16]. M bands are where protein with the wild type genotype is found (M is glycosylated and so wild type is used in reference to the non-glycosylated form). Proteins in the S band are mildly polymerogenic and are of high abundance in southern European populations [17], and it is rare for carriers to show symptoms [18]. The homozygous Z allele is responsible for severe α_1-antitrypsin deficiency and is caused by the mutation E342K, which is positioned at the top of β-sheet A where it disrupts a salt bridge (E342 - K290), thereby disrupting the structure of the protein and promoting polymerisation. The mutation is carried heterozygotically by approximately 4 % of people of Northern European descent and 1 in 2,000 are homozygous [14].

6.1.3 Treatment of severe α_1-antitrypsin deficiency

10-13 % of patient mortality from severe α_1-antitrypsin deficiency is due to liver cirrhosis [12, 17]. Currently there is no treatment for the build up of polymers in the liver. When the liver fails, patients undergo liver transplantation with a 5 year survival rate of 83 % [19]. As the liver is taken from a patient who does not carry the Z allele, the treatment cures α_1-antitrypsin deficiency as the transplanted liver produces M α_1-antitrypsin giving adequate protection to lung tissue [19].

Respiratory failure is responsible for 50-72 % of deaths caused by α_1-antitrypsin deficiency [12, 17, 20]. This issue normally initially manifests as chronic obstructive pulmonary disease (COPD), and treatment for this is the same as with non-α_1-antitrypsin related disease, such as influenza vaccinations, bronchodilators and oxygen administration [21, 22]. These therapeutic strategies however, do not treat the underlying cause that is lack of functioning α_1-antitrypsin.

A direct treatment for COPD caused by α_1-antitrypsin deficiency involves transfusion of purified and pooled human plasma α_1-antitrypsin. This was suggested as a treatment in 1981 [23], and was approved by the United States Food and Drug Administration (FDA) for use as treatment in 1987 [24]. The effectiveness of this treatment is questionable and the governments of many other countries, including the UK, have still not granted a licence [25].

The augmentation therapy is administered weekly as a 60 mg/kg infusion [26]. This is said to increase the blood concentration to a protective threshold of 11 μmol/L which prevents emphysema by adequately controlling levels of neutrophil elastase [27]. This replacement therapy does not however remedy the build up of polymers in hepatic cells and the associated cirrhosis.

6.1.4 Evolutionary benefit of α_1-antitrypsin deficiency

The Z mutation is of higher prevalence than would happen by chance [28] and this indicates that there is likely an evolutionary benefit to the allele. It has

been shown that patients carrying the Z allele display an increased inflammatory response [28]. This could reduce the mortality risk from infections, and thereby increase the life span of carriers.

Inflammation is an endogenous response to infection and so causes upregulation of acute-phase proteins such as α_1-antitrypsin. The liver produces more protein, which increases the concentration of α_1-antitrypsin at the site of inflammation. This effect is amplified as the lungs [29] and gut [30] can both produce α_1-antitrypsin and are the two most common sites for pathogenic invasion. The inflammation response also causes an increase in temperature and a lowering of pH. These two factors in combination with the increase in α_1-antitrypsin concentration cause an increase in the propensity for Z mutant α_1-antitrypsin to polymerise [31], resulting in polymer formation at the site of infection [32]. These polymers are a chemoattractant for cytokines and other inflammatory response proteins creating a positive feedback loop and amplifying inflammation [33, 34].

The increased inflammatory response leads to a lower probability of infections spreading. In 1928, Alexander Fleming discovered penicillin, prior to this the most common cause of death was due to infections. In 1900, 40 % of deaths in the USA were caused by gastroenteritis, tuberculosis, pneumonia and influenza [28], and so reduced risk of infection would be a highly beneficial trait.

The deleterious nature of α_1-antitrypsin deficiency has been amplified due to the dramatic reduction in mortality caused by infection, and it wasn't until the 1960's that α_1-antitrypsin deficiency was discovered and linked to emphysema and liver disease [8, 9].

During the 19^{th} century the average life expectancy was 43 years in the USA [28], thereby precluding the risk factor of contracting emphysema (median age for detection of emphysema in ZZ patients is 53 for non-smokers [12]) as half the population would not live to the age where this risk factor becomes significant. A large exacerbating factor for α_1-antitrypsin deficiency derived emphysema is tobacco smoking which was also a recent introduction in the context of evolutionary time scales [35].

Infant mortality during the 19th century was over 10 % in the UK [36], whereas the risk of mortality during childhood due to ZZ α_1-antitrypsin deficiency induced liver disease is only 1-2 % [28].

These factors combined show that in the pre-antibiotic era, the Z mutation could be of more benefit than harm, explaining the high prevalence of this allele.

6.1.5 Loop-sheet polymerisation model

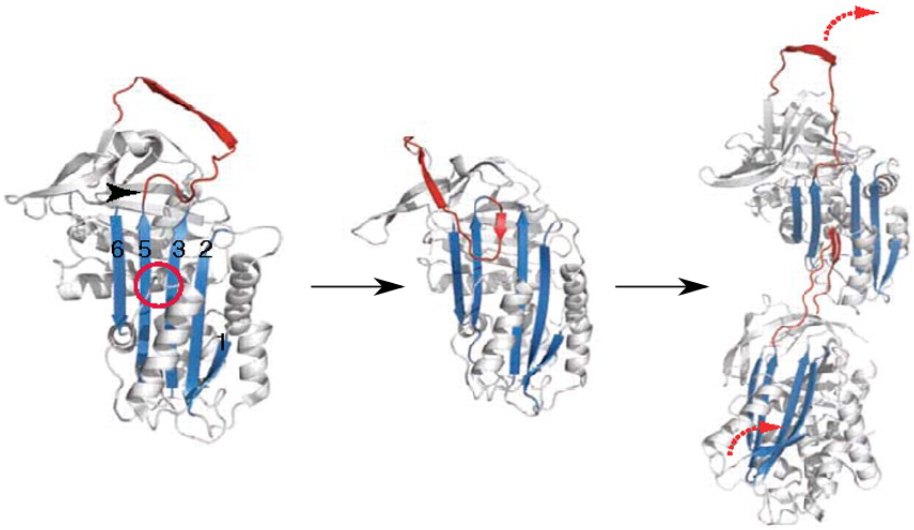

Figure 6.2: The α_1-antitrypsin loop-sheet model of polymerisation. The reactive centre loop (RCL) is coloured red and β-sheet A is coloured blue. (left) The native conformation of α_1-antitrypsin, the position of the Z mutation is shown by the arrow head, and the numbering of the strands of β-sheet A are also shown. (middle) Pre-polymerisation intermediate conformation of α_1-antitrypsin, the RCL is partially inserted into β-sheet A. (right) The classical model of polymerisation showing the interaction between two α_1-antitrypsin molecules. Figure adapted from [37].

The 'classical' model for polymerisation involves the RCL (loop) of one molecule inserting into the centre of β-sheet A (sheet) of another, between β-strands s4A and s5A, mimicking the native interaction with a protease and consequently creating highly stable polymers (Figure 6.2). As this interaction leaves the RCL of the second molecule exposed, it is able to bind into the β-sheet of the next protein allowing for linear chains of polymers as seen

in electron microscopy images [38]. The Z mutation destabilises β-sheet A of the molecule facilitating the occupation of an intermediate conformation (M*) [38] and allowing the insertion of the reactive centre loop [31, 39].

The validity of this model was supported by an X-ray crystallography structure of α_1-antichymotrypsin in an intermediate M* state similar to the α_1-antitrypsin conformation shown in Figure 6.2 (centre). The protein was the mutant form Leu-55-Pro, which is known to polymerise [40]. This model was uncontested until a crystal structure of a serpin polymer was published in 2008 [41].

6.1.6 β hairpin polymerisation model

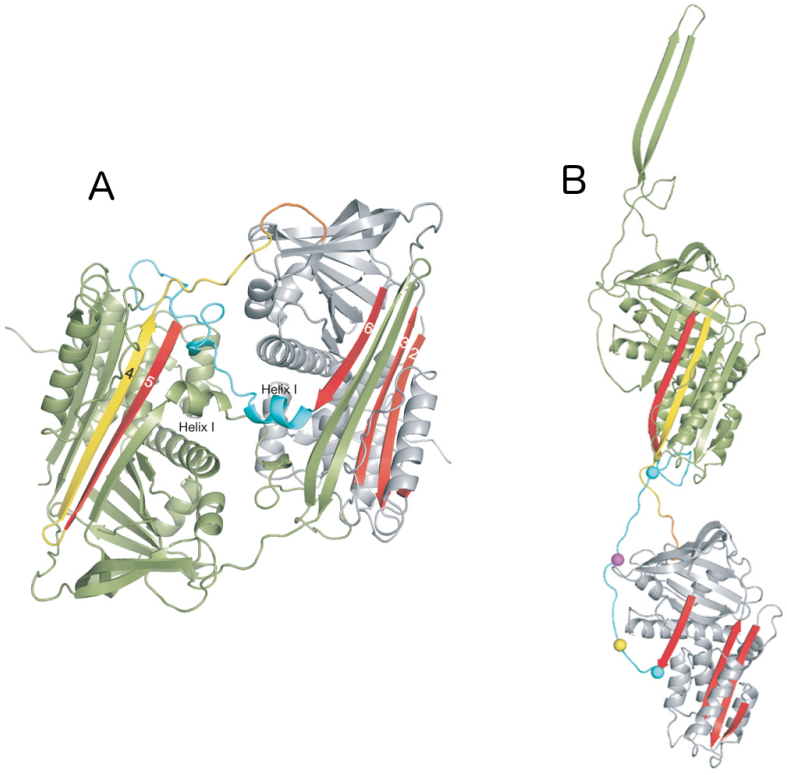

Figure 6.3: β hairpin model of serpin polymerisation. Antithrombin molecules are coloured grey and green, the grey models have β-sheet A coloured red and the reactive centre loop coloured yellow. Shown are a X-ray crystallography structure of a self-terminating antithrombin dimer (A), and a model for how linear polymers could be created with this linkage. Figure adapted from [41].

The β hairpin polymerisation model involves a more extensive domain swap

than in the loop-sheet model. α-helix I is unfolded which allows β-strand s5A to form a β-hairpin with the RCL. This protruding hairpin can then insert into β-sheet A of another molecule by displacing s5A [41], as shown in Figure 6.3.

The polymerisation model was extended to α_1-antitrypsin from an X-ray crystallography structure of a different serpin, antithrombin. The structure was of a wild type antithrombin dimer, and polymerisation was achieved using low pH as a denaturant. The dimer was self-terminating (Figure 6.3A), but a model was also produced to show how linear polymers could form (Figure 6.3B).

2C1 is a monoclonal antibody which was raised to bind to Z mutant *ex vivo* polymers which were extracted from the livers of patients. It has been shown that this antibody also binds to M α_1-antitrypsin polymers which are created by heating the protein [37, 42]. A study in 2009, largely discredited the β hairpin polymerisation model for α_1-antitrypsin as these polymers did not bind 2C1. An additional ion mobility mass spectrometry analysis was conducted where heated M polymer CCS was compared to molecular models of the loop-sheet and β hairpin. As the M protein is glycosylated and the models were void of glycosylates, the experiment compared the percentage change in CCS from monomer to dimer rather than the absolute CCS values. The results showed that the percentage increase in CCS for the experimental data (174 %) was close to that of the loop-sheet model (176 %), but the increase was substantially larger for the β hairpin model (225 %) [37].

6.1.7 C-terminal domain swap polymerisation model

Following the β hairpin model, Yamasaki *et al.* published a crystal structure demonstrating a new mechanism for α_1-antitrypsin polymerisation consisting of a C-terminal domain swap [43]. In this model the β-strands s4B, s5B and s1C which are on the back of the protein in relation to β-sheet A are not folded as in the native protein structure. Instead β-strands s4B and s5B form a β hairpin which then fill their equivalent positions in another α_1-antitrypsin molecule (Figure 6.4). The increase in the stability of the polymer

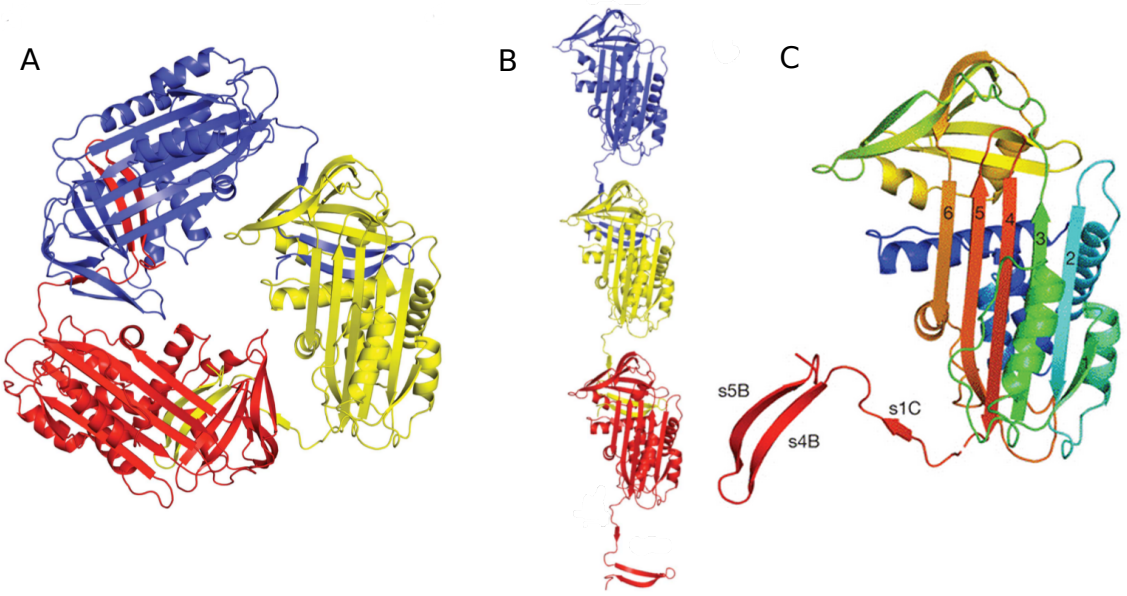

Figure 6.4: X-ray crystallography structure of a self-terminating α_1-antitrypsin trimer displaying the C-terminal domain swap model (A). Additionally shown is a model for how linearised polymers could form (B). (C) Labelled structure, coloured from N-C terminus as blue to red. Strands in β-sheet A are numbered. Figure adapted from [43].

occurs through insertion of the RCL into the centre of β-sheet A, mimicking the interaction with proteases. It is proposed that this fold would not occur spontaneously, rather polymers would occur during protein folding, with the coupling being made before β-sheet C had folded.

In order to create the X-ray crystallography structure, the researchers mutated residues on β-strands s5A (cysteine 292) and s6A (cysteine 339), to allow the formation of a disulphide bond holding these two strands together. This created short polymers which were resistant to further elongation and were able to be crystallised. These polymers were additionally found to bind to the 2C1 antibody, supporting the model as the physiological polymer structure [43].

6.1.8 Peptides block polymerisation

Blocking the polymerisation of α_1-antitrypsin would alleviate the pathogenic effects within liver cells. For this reason small molecules [45] and peptide ho-

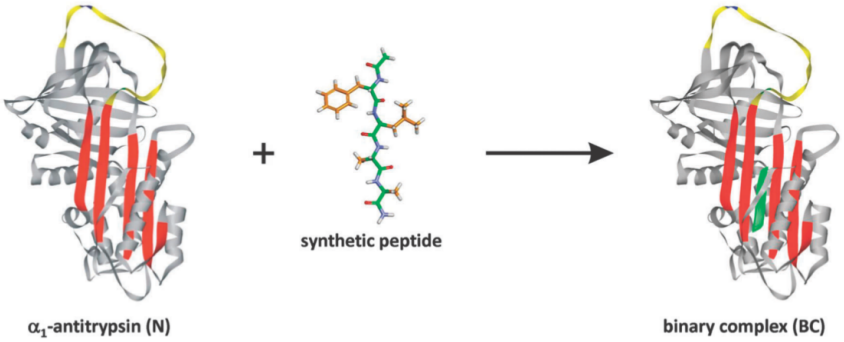

Figure 6.5: Proposed model for Ac-TTAI-NH_2 binding with α_1-antitrypsin. Figure reproduced from [44].

mologs of the RCL [46] have been produced which target the RCL insertion point for polymerisation. Previously the shortest peptide shown to block polymerisation was a 6-mer [47]. Following this a tetrapeptide was found which had similar efficacy to the 6-mer [48], the importance of reduction in peptide length is due to the increased pharmacological significance as a potential drug molecule. The same group in 2009 used a combinatorial chemistry approach and the tetrapeptide Ac-TTAI-NH_2 was identified [49]. This molecule had improved specificity and it was shown that a 2hr incubation in a 10:1 (peptide:Z α_1-antitrypsin) molar ratio was sufficient to achieve the binary complex as probably the most abundant species. It was additionally shown that under these conditions no M-peptide complex was detectable by polyacrylamide gel electrophoresis (PAGE) [49]. The group has proposed a mechanism of binding as shown in Figure 6.5, where a single Ac-TTAI-NH_2 molecule binds in the centre of β-sheet A through a similar mechanism as with the native interaction with a protease increasing the thermal stability of the molecule [44].

6.1.9 Aims

This chapter aims to investigate the interaction between α_1-antitrypsin and Ac-TTAI-NH_2 using the tools introduced in this thesis. These studies will then be able to provide further information towards the feasibility of testing putative drug molecules using wild type, M and Z α_1-antitrypsin as well as in a high throughput environment. Finally, the different polymerisation

mechanisms will be investigated with Z polymers extracted from human liver using ion mobility mass spectrometry.

6.2 Methods

6.2.1 Sample sources

β-lactoglobulin from bovine milk and concanavalin A from *Canavalia ensiformis* were purchased from Sigma Aldrich (St. Louis, MO). Ac-TTAI-NH_2 was ordered from Pepceuticals (Leicester, UK).

6.2.2 Recombinant α_1-antitrypsin expression and purification

The α_1-antitrypsin gene, encoded in cDNA with a 6*HisTag, was ligated into the pQE31 plasmid vector (kindly provided by Dr. Nyon) and transformed into BLT-21 where gene expression is inducible through the addition of isopropyl β-D-1-thiogalactopyranoside (IPTG). Once optical density of the culture at OD_{600} reaches 0.6, protein expression was induced through the addition of 1 mM IPTG and incubated for 8 hours at 30 °C to maximise protein expression. Cells were pelleted through centrifugation and resuspended in a solution of sodium phosphate buffer (10 mM, pH 8.0), NaCl (0.5 M), and imidazole (20 mM) before exposure to a cell disruptor where lysis released the expressed product. Centrifugation was used to remove the insoluble cellular debris and the supernatant containing any soluble protein was loaded directly onto a HisTrap Crude FF column (5 ml; GE Healthcare), then washed and eluted using increasing concentrations of imidizole ranging from 20 mM – 200 mM. In order to concentrate and further purify the protein, the fractions released from the column were subject to dialysis in 10 mM Tris, pH 8.0, 1 mM EDTA, and 1 mM β-mercaptoethanol before a final filtration when loaded onto a a HiTrap-Q Sepharose column (5 ml; GE Healthcare). Incremental concentrations of NaCl from 0-0.5 M were used to elute the his tagged prod-

uct and fractions were collected and combined before another stage of dialysis against 25 mM Na_2HPO_4, 50 mM NaCl and 1 mM EDTA at pH 8.0.

6.2.3 Ac-TTAI-NH_2 α_1-antitrypsin titration

α_1-antitrypsin was buffer exchanged into 250 mM ammonium acetate (pH 7) by 3 rounds of dilution-concentration with 10 kDa Millipore Amicron Ultra centrifuge filters. The protein concentration was then determined using a Qubit 2.0 fluorometer. Two Ac-TTAI-NH_2 stocks were made at 30 μM and 250 μM to minimise error from pipetting small volumes. The α_1-antitrypsin was aliquoted, peptide was added at relevant volume and the samples were diluted with 250 μM ammonium acetate to a final α_1-antitrypsin concentration of 15 μM in 100 μL. Samples were then incubated for 72 hours at 37.5 °C before analysis. The mass spectrometer settings used are shown in Table 6.1.

Setting	Titration	Plasma Z	Oligomers
Capillary Voltage	1.2 kV	1.1 kV	1.0 kV
Sampling Cone	45 V	40 V	50 V
Extraction Cone	1 V	1 V	1 V
Source Temperature	40 °C	40 °C	40 °C
Trap Collision Energy	6 V	6 V	20 V
Transfer Collision Energy	4 V	4 V	10 V
Trap Pressure	1.5x10^{-2} $m$bar	1.5x10^{-2} $m$bar	2.5x10^{-2} $m$bar
Bias	4 V	4 V	4 V
Backing Pressure	0.21 $m$bar	0.21 $m$bar	0.545 $m$bar

Table 6.1: Mass spectrometer settings used for; Ac-TTAI-NH_2 α_1-antitrypsin titration experiments, Z α_1-antitrypsin extracted from plasma, and M and Z oligomer experiments.

6.2.4 Ac-TTAI-NH_2 α_1-antitrypsin unfolding experiments

α_1-antitrypsin dissolved in 25 mM Na_2HPO_4, 50 mM NaCl and 1 mM EDTA at pH 8.0 at a concentration of 1.5 mg/ml was incubated with Ac-TTAI-NH_2 in 20 mM ammonium acetate at a 20:1 peptide to protein molar ratio for

20 hours. The protein was then buffer exchanged into 150 mM ammonium acetate (pH 7.4) using a BioRad BioSpin column followed by 2 rounds of dilution-concentration with 10 kDa cut-off Millipore Amicron Ultra filters.

Samples were then analysed after quadrupole isolation of the +13 charge state of the apo and double bound forms of the protein. The ion mobility mass spectrometry calibration curve was determined using β-lactoglobulin monomer and dimer as well as concanavalin A tetramer. The mass spectrometer settings used are shown in Table 6.2.

Setting	Unfolding	Z Oligomers
Capillary Voltage	1.1 kV	1.0 kV
Sampling Cone	30 V	50 V
Extraction Cone	1 V	1 V
Source Temperature	40 °C	40 °C
Trap Collision Energy	Variable	20 V
Transfer Collision Energy	10 V	10 V
Trap Pressure	5 x10^{-2} $m$bar	5 x10^{-2} $m$bar
Mass Range	1,000-8,000 m/z	1,000-16,000 m/z
Bias Voltage	20 V	20 V
Backing Pressure	0.42 $m$bar	0.545 $m$bar
IM Wave Height	10 V	10 V
IM Wave Velocity	350 $m \cdot s^{-1}$	350 $m \cdot s^{-1}$
Transfer Wave Height	8 V	8 V
Transfer Wave Velocity	150 $m \cdot s^{-1}$	100 $m \cdot s^{-1}$
IMS pressure	5.20x10^{-1} $m$bar	5.20x10^{-1} $m$bar
LM/HM Resolution	5/15	15/5

Table 6.2: Ion mobility mass spectrometry settings used for Ac-TTAI-NH_2 α_1-antitrypsin gas-phase unfolding experiments and oligomeric Z α_1-antitrypsin extracted from hepatocytes.

6.2.5 M polymer preparation

Monomeric M protein was kindly donated by Dr. Imran Haq who extracted the protein from healthy human patient plasma. Polymers were created *in vitro* by incubating 0.5 mg/ml purified protein solutions under various conditions. Heat induced polymers were incubated in a pH 7.4 solution of 10 mM Na_2PO_4, 100 mM NaCl and 0.02 % NaN_3 at 50 °C for 48 hours. Acid

induced polymers were formed by incubating at 25 °C for 7 days in a 0.1 M solution of sodium acetate at pH 4.8. Urea denaturing induced polymers were created by incubating for 14 days at 25 °C, in pH 8.0 solution of 40 mM Tris containing 4 M urea. Guanidine hydrochloride (GuHCL) was also used to create polymers, the protein was incubated for 11 days at 25 °C in pH 8.0 Tris in addition to 3 M GuHCL.

The polymers were then fractionated using size exclusion chromatography, and the fractions were assessed for degree of polymerisation with polyacrylamide gel electrophoresis. Fractions with high levels of polymerisation were selected, pooled and dialysed into 5 litres of pH 7 150 mM ammonium acetate for 48 hours.

Before MS analysis the samples were concentrated, diluted and concentrated again with Millipore centrifuge tubes and pH 7 150 mM ammonium acetate.

6.2.6 Glycosylated ion mobility Z α_1-antitrypsin experiments

Z mutant *ex vivo* polymers were provided by Sarah Faull. In brief the liver tissue was removed from disease state liver after a patient transplant. After filtration to remove fibrous tissue, the inclusion bodies were separated using sucrose gradient fractionation. The inclusion bodies were ruptured using sonication, releasing the soluble α_1-antitrypsin polymers. Finally ultracentrifugation was used to separate insoluble protein and debris, with the protein removed as supernatant. The protocol is described in greater detail elsewhere [50, 51].

To prepare the sample for mass spectrometry analysis, the protein was first dialysed into 5 litres of 200 mM ammonium acetate at pH 7 for 48 hours. The sample was then filtered with Corning Costar (NY, USA) Spin-X 0.22 μM centrifuge tubes to remove debris before an additional step of buffer exchange using a BioRad BioSpin tube, and finally concentrated using a 50 kDa cut off Millipore Amicron Ultra centrifuge tube. The ion mobility calibrants used in

the analysis were monomeric and dimeric β-lactoglobulin and concanavalin A tetramer, and the mass spectrometer settings are shown in Table 6.2.

6.3 Results and discussion

6.3.1 Ac-TTAI-NH_2 titration with wild type α_1-antitrypsin

The interaction between Ac-TTAI-NH_2 and the Z mutant is thought to mimic the interaction with the reactive centre loop (RCL) of α_1-antitrypsin, thereby stabilising the protein as in its native interaction with proteases. Previous studies have found that Ac-TTAI-NH_2 does not bind to wild type α_1-antitrypsin [48, 49], presented here is evidence to the contrary.

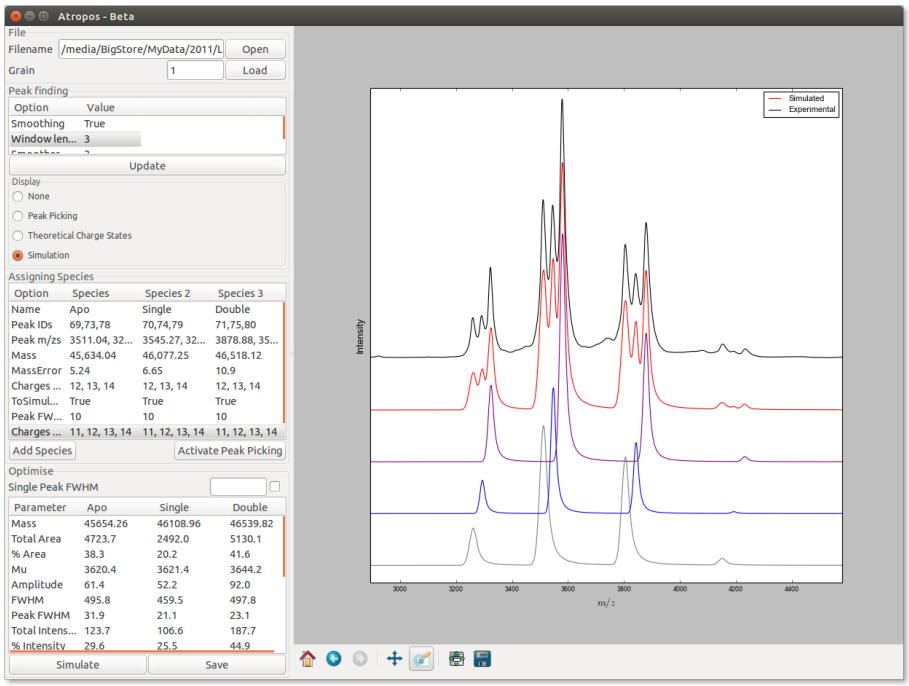

Figure 6.6: Deconvolution of the 0.3:1 peptide to protein mass spectrum using the Amphitrite Atropos graphical user interface. The individual deconvoluted components are shown; apo (grey), single bound (blue) and double bound (purple). The sum of the simulated parts is shown in red and the original experimental data are black.

Figure 6.6 shows a mass spectrum of wild type α_1-antitrypsin after being incubated for 3 days at 37 °C with Ac-TTAI-NH_2. The wild type protein has bound to the Ac-TTAI-NH_2 with a stoichiometry of 2:1 (ligand:protein). This means that the current model for binding which involves a single Ac-TTAI-NH_2 molecule should likely be revised if the Z mutant binds similarly. The finding could further support the hypothesis that the interaction mimics the

interaction with the RCL.

During the interaction with a protease, a large groove is created and filled by the RCL, which is evidenced by the X-ray crystallography structure in Figure 6.7 (1EZX.pdb). The reactive centre loop insertion is coloured alternately, per residue, in red and orange and this representation shows that there are 11 RCL residues in the cleft. The finding that two 4 residue Ac-TTAI-NH_2 molecules bind to α_1-antitrypsin, as shown in Figure 6.6, further supports the hypothesis that the molecules are binding in the cleft and mimicking the protease interaction.

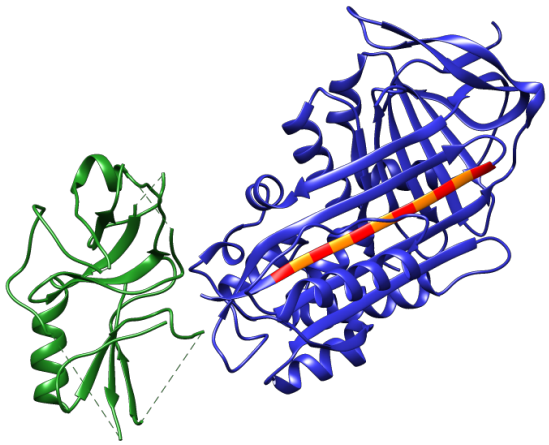

Figure 6.7: Crystal structure of α_1-antitrypsin (blue) in complex with a protease, trypsin (green). Additionally the reactive centre loop has been coloured red and orange with the colour alternating per residue (1EZX.pdb) [3].

These data suggests that the interaction is cooperative as the double bound form is of higher abundance than the single bound form. To investigate this further a titration was performed with varying Ac-TTAI-NH_2 concentrations and the results are shown in Figure 6.8.

The titration data shows that the single bound species is evidently less abundant than the double bound species at Ac-TTAI-NH_2 concentrations of 0.5:1 and above, further confirming the cooperativity of the interaction.

At the highest peptide concentration there is the appearance of an additional peak corresponding to the mass of one α_1-antitrypsin and 3 Ac-TTAI-NH_2 molecules. The spectra were acquired straight after incubation and so

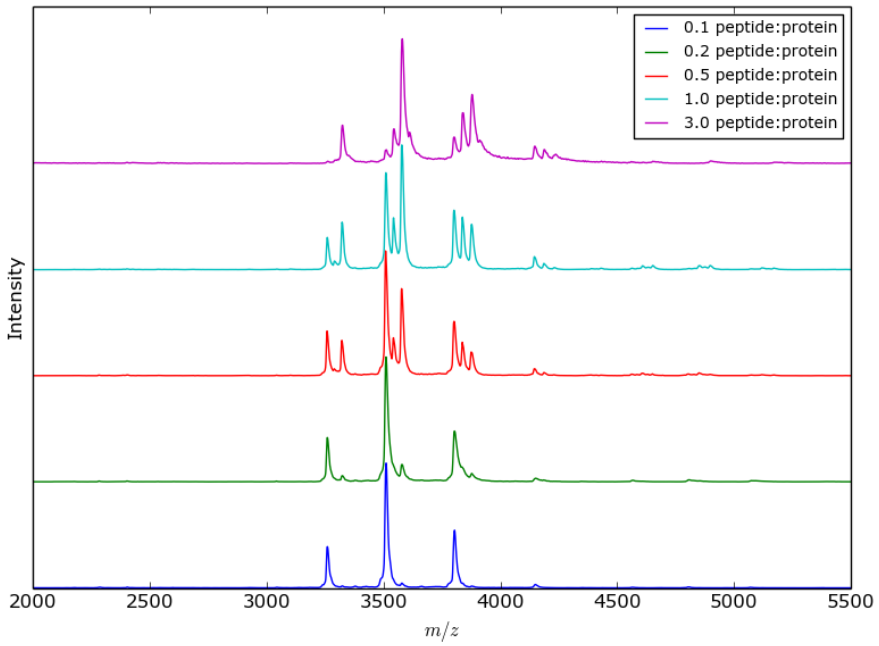

Figure 6.8: Titration mass spectra of Ac-TTAI-NH_2 and α_1-antitrypsin.

there was still a high peptide concentration in the buffer, potentially suggesting that this additional interaction is non-specific. To reduce the contribution of non-specific binding, after incubation the sample was concentrated and resuspended several times using a 10 kDa molecular weight cut off filter. This amounts to a 10,000 fold dilution for unbound Ac-TTAI-NH_2 (>500 Da), and so it could be approximated that the final molar ratio of unbound Ac-TTAI-NH_2 to protein would be 0.003:1 and as seen in Appendix 6.24, this indicates that it is highly unlikely that the interaction observed is due to non-specific binding.

The process of dilutions with concentration columns changes the distribution of bound species as it changes the solution equilibrium. For this reason the incubation had to be conducted in a mass spectrometry compatible buffer (ammonium acetate). The interaction between Ac-TTAI-NH_2 and wild type α_1-antitrypsin is very slow (see Appendix 6.23) and during the time taken for the reaction to reach equilibrium, much of the protein precipitates due to the ammonium in the buffer [52]. This means that the dissociation constant (K_d) cannot be calculated for the interaction as the apo monomeric folded protein

concentration ($[P]$) needs to be known for the calculation (Equation 6.1) as well as ligand concentration ($[L]$) and complex concentration ($[PL]$).

$$K_d = \frac{[P] \cdot [L]}{[PL]} \tag{6.1}$$

With the target of creating a framework for analysing new ligands which could have faster reaction rates, it is still useful to assess the implementation of the new methodology towards the determination of species abundance for K_d calculations. Existing examples of estimating abundance for calculating K_d values have used peak heights [53]. The problem with using peak heights is demonstrated in Figure 6.9.

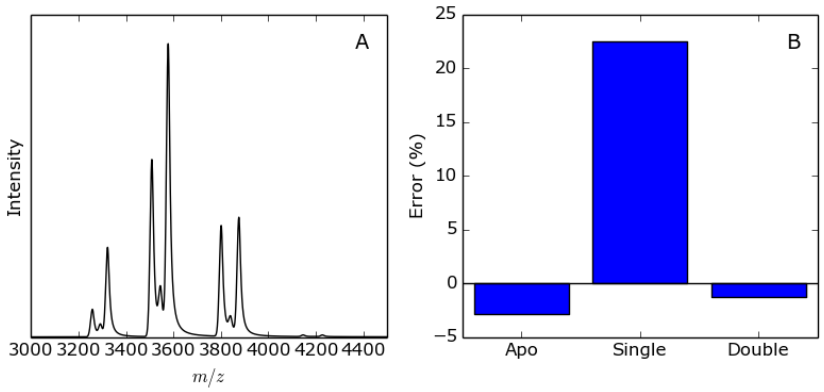

Figure 6.9: Synthetic data based on the 0.3:1 Ac-TTAI-NH_2:α_1-antitrypsin mass spectrum (A). When analysing A using peak heights, this graph shows the percentage error from the proportional abundance of each of the bound states (B).

A native protein mass spectrometry peak consists of a single peak of the unadducted protein as well as several $n \cdot adduct + protein$ peaks which overlap to form the overall peak seen in the mass spectrum. Different charge states have different numbers of adduct peaks as lower charge ions have lower internal energy in the mass spectrometer, which reduces the level of desolvation and adduct stripping. For these reasons, it is correct to use the peak areas instead of peak heights.

The second problem with using peak heights is that protein and ligand peaks can overlap, often resulting in an overrepresentation of low abundance

peaks which occur at the same m/z value as the peak shoulder of a larger peak. To demonstrate this issue a synthetic mass spectrum was created using the masses of α_1-antitrypsin alone and with one and two Ac-TTAI-NH_2 molecules bound (Figure 6.9A). The difference between the areas of the components of the synthetic data and the peak heights are shown in Figure 6.9B. The single bound species is overrepresented by over 20 % as its peak height is increased due to overlap with the apo and double bound species, concomitantly both of the other species are underrepresented.

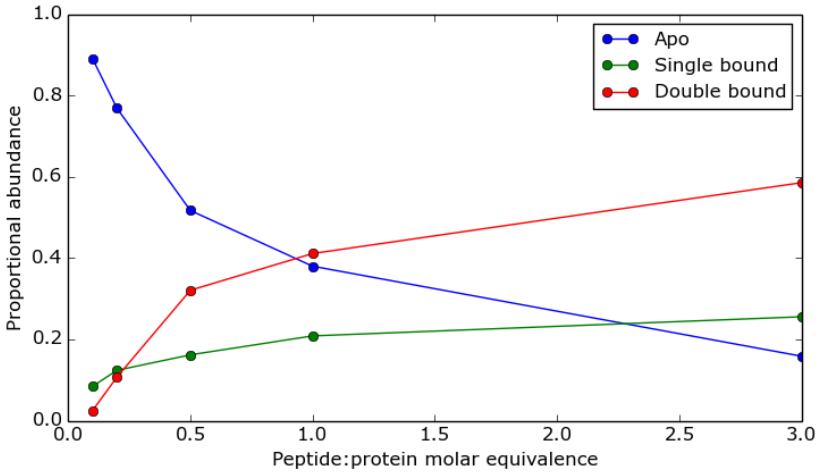

Figure 6.10: Deconvoluted abundances of Ac-TTAI-NH_2, α_1-antitrypsin titration. Results are shown as a proportion of all three states.

The peak areas can be deconvoluted using the mass spectrum fitting algorithm of Amphitrite, and the results are shown in Figure 6.10. The double bound species is more abundant than the single bound after a molar equivalence of 0.2:1 once again confirming the cooperativity of the interaction. Due to the precipitation of the protein during the long incubation time required to reach equilibrium it is not possible to calculate dissociation constants K_d using these data.

6.3.2 Unfolding experiments

Now that it has been determined that wild type α_1-antitrypsin binds to Ac-TTAI-NH_2, it would be beneficial to further characterise the interaction. It is logical that the blocking of aggregation caused by Ac-TTAI-NH_2 binding to Z mutant α_1-antitrypsin, would be as a result of the stabilisation of the protein. This hypothesis is tested here, using gas-phase unfolding experiments and the methodology outlined in Chapter 5. If the technique is able to determine changes in stability caused by Ac-TTAI-NH_2, it will be possible for the finding to be used in future studies for screening putative drug molecules which block α_1-antitrypsin aggregation.

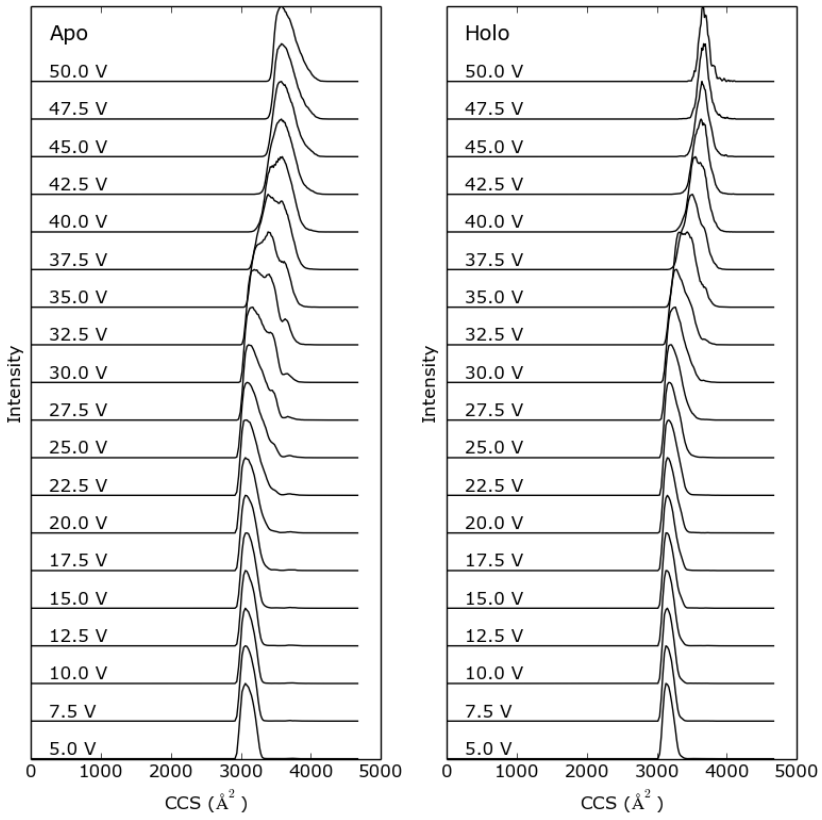

Figure 6.11: Stacked plot of arrival time distributions (calibrated to CCS), for apo and holo α_1-antitrypsin.

Existing data analysis techniques

After completing the automated unfolding experiment as described previously, the relevant information from the MassLynx raw files is extracted and

converted to CCS using the Amphitrite software package. The CCSDs (calibrated ATDs) are displayed in Figure 6.11, the apo distributions appear to generally be wider than for the holo protein, indicating Ac-TTAI-NH_2 binding results in a reduction in conformational variability as a result of binding. There also appears to be additional partially resolved peaks in the apo data, further confirming the reduction in conformational variability.

These observations are generally self-evident from the graphs; however, they are not quantitative and so it would be difficult to screen many putative ligands, as it would require an operator to manually check the data. It is also non-trivial to assess the extent of the unfolding from these plots, and additional data analysis would be required to elucidate any changes.

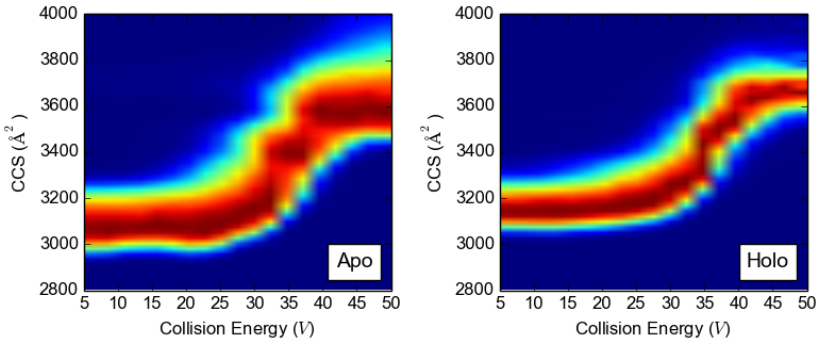

Figure 6.12: Collision induced unfolding (CIU) fingerprint analysis of apo and holo α_1-antitrypsin gas-phase unfolding.

CIU fingerprint analysis was also carried out on the unfolding experiment data, and is shown in Figure 6.12. Similarly to the stacked CCSD data representation (Figure 6.11) it is axiomatic that the conformational variability is greater for the apo protein. This is shown as the increased width of the distributions along the CCS axis. Using this method however has lost the information regarding the additional partially resolved peaks in apo α_1-antitrypsin CCSDs. Observing the difference between low and high energy regions for each protein, it can be seen that the difference in peak top CCS is greater for the Apo. This indicates that the interaction with Ac-TTAI-NH_2 reduces the amount the protein unfolds as well as reducing conformational variability.

As with the stacked CCSDs, this method is not quantitative and so it would require a trained human to manually check whether a putative ligand was

affecting the conformational stability and variability and so is not appropriate for automated analysis.

Summary statistics

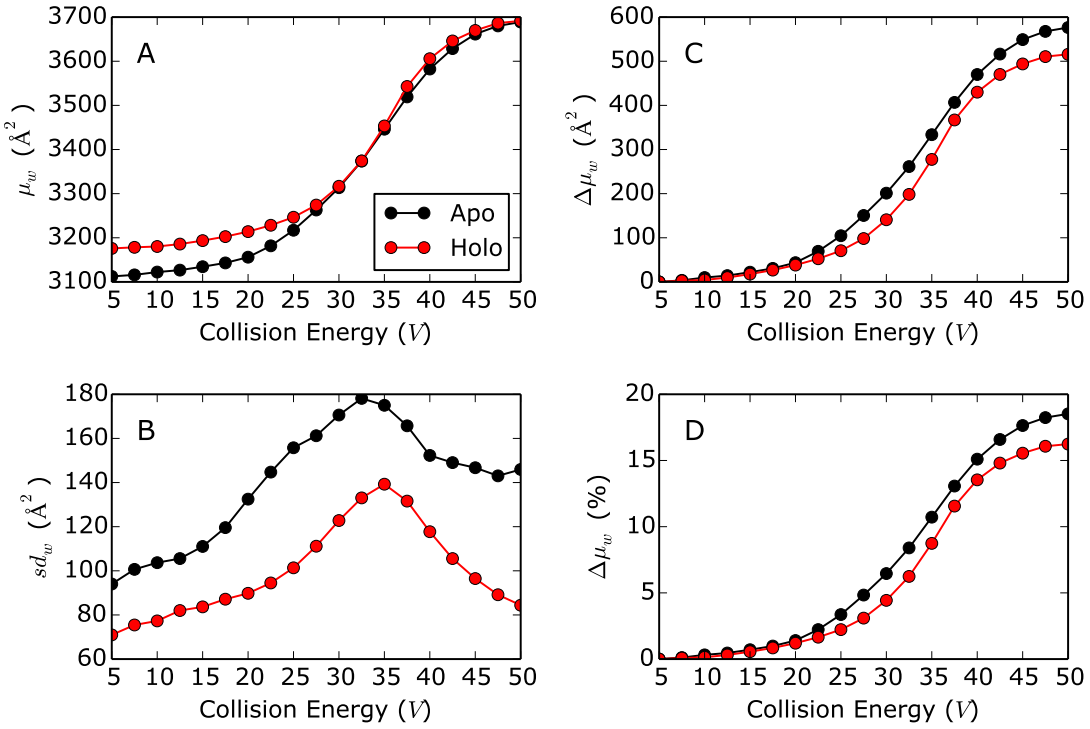

Figure 6.13: Summary statistic analysis of apo and holo α_1-antitrypsin. Unfolding curves (A), variability curves (B), change in CCS from lowest collision energy represented as an absolute value (C) and as percentages (D).

The summary statistic analysis as introduced in the Chapter 5 was performed on the data and the results are shown in Figure 6.13. Looking at the average cross section of the bulk protein (Figure 6.13A) shows that the average cross section is higher at low energy (5 V) for the holo form of the protein (3,176 Å^2) versus the apo form (3,112 Å^2). Contrastingly, as the collision energy increases the average CCS of the apo protein increases more readily with similar values observed at the highest energies. To normalise for the higher initial starting CCS of the holo protein, the difference (Figure 6.13C) or percentage difference (Figure 6.13D) in CCS from lowest collision energy can be analysed. In both forms of the analysis, the change in CCS is less for the holo form throughout the unfolding experiment, with the effect more pronounced

in the percentage difference plot. This shows that the ligand interaction stabilises the protein and reduces the extent to which it unfolds. The variability curve is shown in Figure 6.13B, throughout the experiment the holo protein has a lower standard deviation indicating reduced conformational variability.

As all of these results give numerical values, they can be represented in tables as shown in Table 6.3 and 6.4. When screening putative drug molecules, threshold values could be given for whether the molecule should be brought forward for manual analysis. An example of this would be to select ligands where at 30 V the percentage change in CCS is under 5 % and the standard deviation is under 150 Å^2.

	30 V	50 V
Apo ΔCCS	6.46 %	18.52 %
Holo ΔCCS	4.44 %	16.24 %

Table 6.3: Table showing the change in CCS from the most native-like analysis to a given collision energy as a percentage.

	5 V	30 V	50 V
Apo σ CCS	94.02 Å^2	170.60 Å^2	145.93 Å^2
Holo σ CCS	70.95 Å^2	122.79 Å^2	84.41 Å^2

Table 6.4: Table showing the weight standard deviation values of apo and holo α_1-antitrypsin at 3 collision energies.

Multicomponent analysis

The confirmation that the Ac-TTAI-NH_2 interaction stabilises α_1-antitrypsin has, so far, been observed on a bulk level. To understand what is happening with individual conformations, the data was deconvoluted using the Chalenger algorithm. The results are shown in Figure 6.14 and the CCS value of the centre of each conformation is shown in Table 6.5. At low collision energies, the holo protein predominantly occupies a single conformation, whereas the apo protein shows a high abundance of two conformations. This would explain why the conformational variability is higher for the apo protein at low energies (Table 6.4). At the two highest collision energies there are two similar abundance conformations present for the apo protein whereas there is only

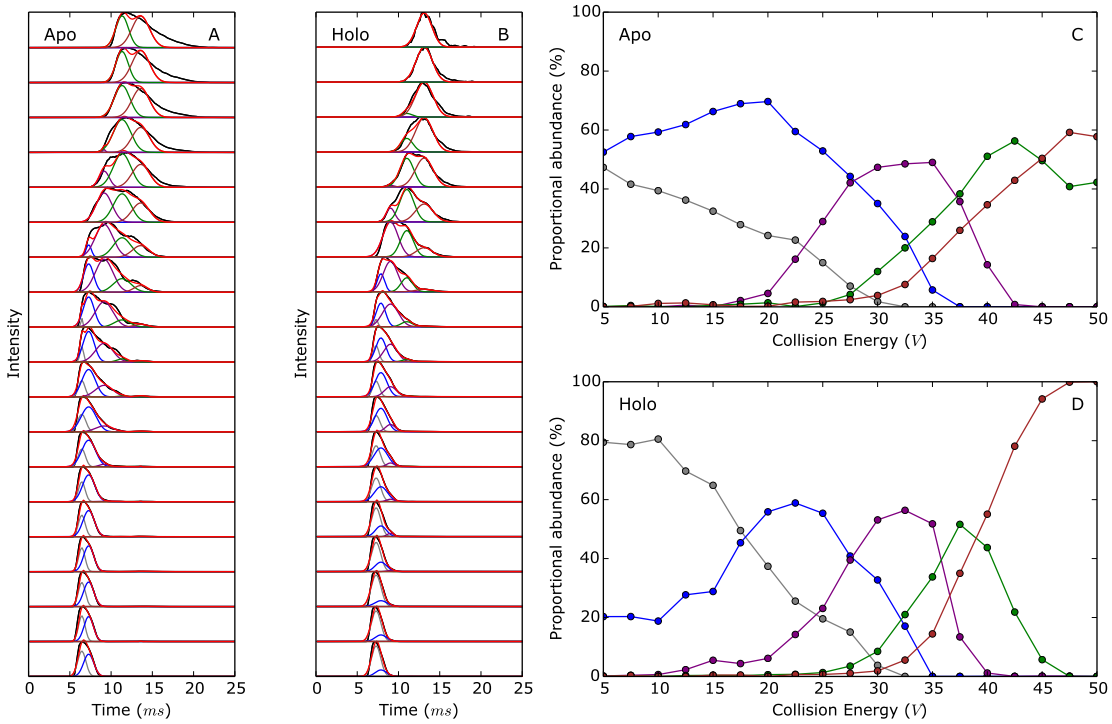

Figure 6.14: Deconvolution of gas-phase unfolding data for apo (A) and double bound (B) α_1-antitrypsin. Experimental data are shown in black, with the sum of the deconvolution in red. The deconvoluted distribution of each conformation is shown with coloured lines and the mean arrival time for each conformation is shown in Table 6.5. The abundance analysis of each conformational family determined by the deconvolution is shown for the apo (C) and holo (D) protein. Arrival time and CCS values for the centre of each conformation is given in Table 6.5.

one for the holo form. This is likely due to experimental constraints, where if one or more ligands dissociate from the holo complex, it will no longer be registered as it moves out of the quadrupole isolated m/z window. The likely result of this is that the conformation demonstrated by the holo form is the most open conformation the protein can populate without the dissociation of a Ac-TTAI-NH_2 molecule.

These effects are demonstrated in Figure 6.14C and D as well. These data show that the main distinguishing factor between the apo and holo form is that ligand binding stabilises a single conformation at lower collision energies in comparison to the apo protein. This feature could then be used for assessing putative ligands; if a ligand is shown to stabilise the protein in terms of

CCS, reduces the conformational variability and the deconvolution shows that there is initially a single predominant conformation, then it is likely that the ligand will have a similar binding mechanism as Ac-TTAI-NH_2 and will block α_1-antitrypsin aggregation. For completeness an analysis of the change in conformation of dissociated holo protein is included in Appendix 6.5.1.

Colour	Apo (t_d)	Apo (Å^2)	Holo (t_d)	Holo (Å^2)
Grey	6.5	3074.8	7.3	3172.8
Blue	7.3	3171.9	7.9	3234.9
Purple	9.2	3358.9	9.0	3347.8
Green	11.3	3538.8	11.0	3514.2
Brown	13.6	3696.5	13.1	3665.5

Table 6.5: Collision cross section values of α_1-antitrypsin conformations, as determined using the Challenger deconvolution algorithm.

6.3.3 Analysis of *ex vivo* α_1-antitrypsin

We were able to obtain small amounts of Z mutant α_1-antitrypsin extracted from human plasma, in monomeric form, as well as extracted from hepatocyte inclusion bodies as oligomers. Various tests for analysing the feasibility of monitoring drug molecule + Z α_1-antitrypsin using native MS and IM-MS are presented here.

Plasma Z α_1-antitrypsin

Homozygous Z mutant α_1-antitrypsin patients still have some protein in their circulation. Some of this protein was extracted from a patient's plasma, and the mass spectrum acquired is shown and deconvoluted in Figure 6.15. Each charge state is made up of multiple peaks owing to glycosylation. The mass as calculated from sequence is 46,589.40 Da (mutation added to UniProt entry P01009), showing that the protein without glycosylates is either not present or at extremely low abundance. Glycosylation can cause problems with native mass spectrometry protein-ligand analysis due to overcrowding of spectra. In an effort to assess whether this would be a problem, a simulation was performed of a mass spectrum with Ac-TTAI-NH_2 and the glycosylated

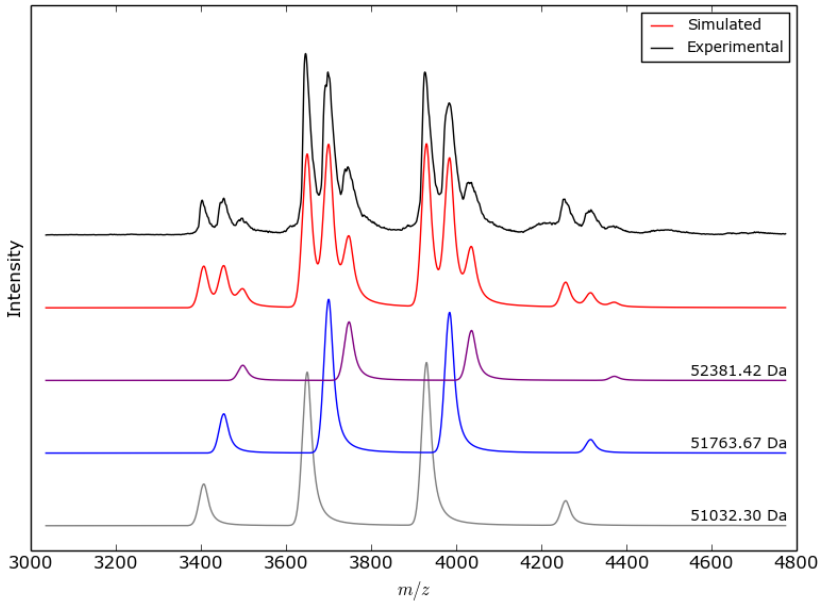

Figure 6.15: Mass spectrum of plasma Z mutant α_1-antitrypsin, deconvoluted using Amphitrite. Grey, blue and purple traces are of different glycosylation states, of monomeric protein.

Z mutant (Figure 6.16).

The simulation was based on the 0.5:1 peptide to protein mass spectrum shown earlier (Figure 6.8). The peak FWHM values were taken from the Z deconvolution in Figure 6.15, with the single and double bound peaks using the same peak FWHM as the unbound form of the same glycosylated species, and the charge state Gaussian FWHM was similarly calculated. The mass was determined by adding the mass of one or two Ac-TTAI-NH_2 molecules to the mass determined by the Z mutant deconvolution. The charge state distribution Gaussian centre was calculated using the m/z difference between differing bound species from the titration deconvolution and adding it to the Z mutant deconvolution result.

Examining the result of the simulation shows a highly congested spectrum, however it would be possible to extract ion mobility data or quadrupole isolate the double bound form of the protein (shaded pink area). This could allow for analysis as shown in Section 6.3.2 to assess the stabilising effect of Ac-TTAI-NH_2 on the Z mutant protein.

When searching for new ligands they may have less mass or only bind

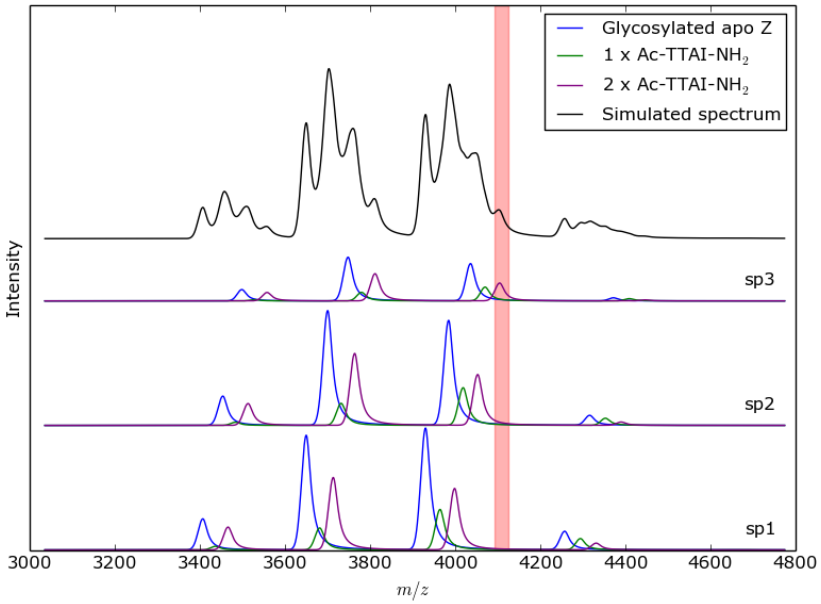

Figure 6.16: Simulation of what a mass spectrum of plasma Z α_1-antitrypsin with Ac-TTAI-NH_2 could look like. The three glycosylated molecular species are labelled sp1-3, and a potential extraction window for ion mobility analysis is highlighted in pink.

with a 1:1 stoichiometry, this would make it more difficult to separate bound species with a quadrupole. The simulation is however, a worst-case example. If the 3:1 mass spectrum was used instead, the bound forms of the protein would be vastly more abundant allowing for analysis of the bound species, with a control sample without ligand being used as the apo analyte.

The work here shows that it is likely that analysis of new ligands identified by interactions with wild type α_1-antitrypsin, would be able to be similarly analysed with the much more scarce Z mutant protein.

Analysing oligomeric glycosylated α_1-antitrypsin

The mass spectra of glycosylated polymeric protein are likely to be highly congested. To test the feasibility of native MS analysis of *ex vivo* polymers of Z mutant α_1-antitrypsin, glycosylated wild type α_1-antitrypsin (M) oligomers were analysed (Figure 6.17). These analyses also indicate whether it will be possible to analyse small molecule binding with this technique.

PAGE analysis has shown that α_1-antitrypsin polymers can be created

using various denaturing conditions [37]. Four of these conditions are tested here and oligomeric species are identifiable up to at least trimer in all cases.

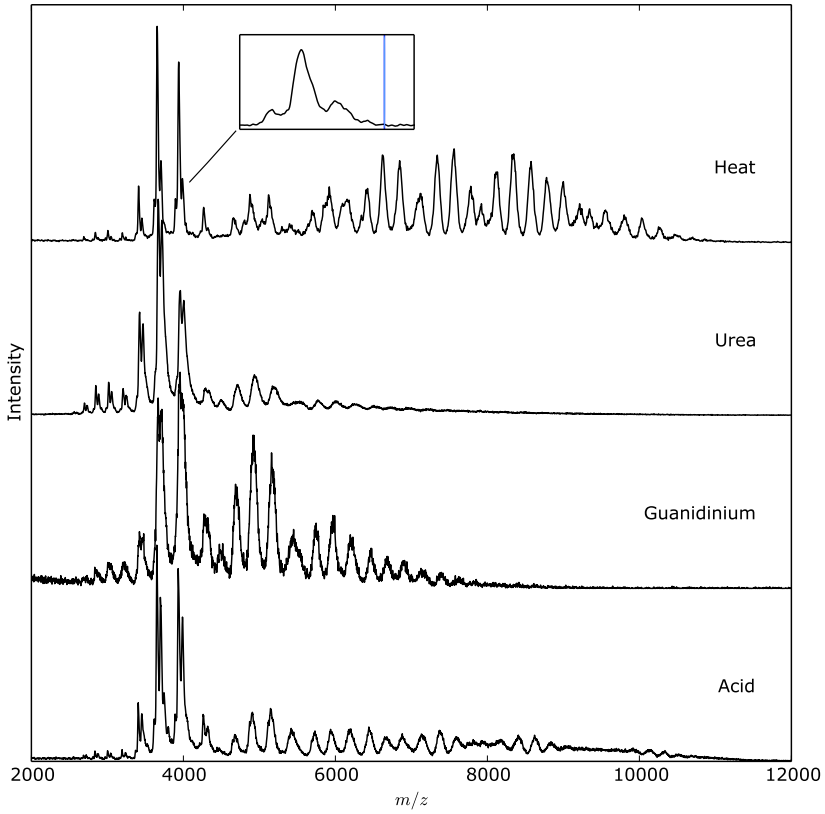

Figure 6.17: Mass spectra of M (glycosylated wild type) α_1-antitrypsin after incubating at different polymerising conditions. The inset shows the +13 charge state peak under heating conditions with the blue line indicating the theoretical m/z value of the largest extent of glycosylation on the monomer in addition to two Ac-TTAI-NH_2 molecules.

Individual peaks are well separated and the m/z ratio of the largest glycosylated species of the heated polymer bound to two Ac-TTAI-NH_2 molecules is shown in the Figure 6.17 inset. The heated polymers are the most relevant for putative drug screening experiments as heating is the only method for creating polymers (without mutating the protein), which are detected by the 2C1 antibody [37]. The m/z region where the double bound protein peak would appear (blue line) does not overlap with other peaks and so unfolding and ion mobility experiments would be straight forward.

The success of the M polymer experiments facilitated the acquisition of a small amount of Z α_1-antitrypsin oligomers, extracted from human hepatocyte

endoplasmic reticulum inclusion bodies.

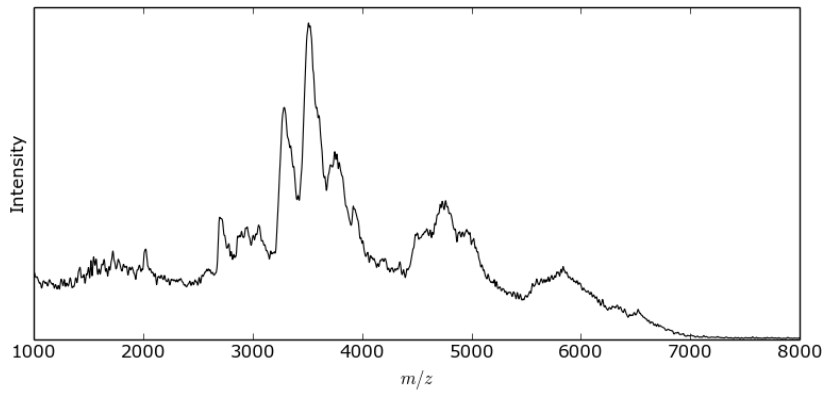

Figure 6.18: Mass spectrum of Z mutant α_1-antitrypsin polymer extracted from human liver.

The sample was not pure with large amounts of residue left after centrifugation which could be lipids. Several rounds of optimisation were performed including: adding dialysis to protein preparation, adding low methanol concentrations (5%), and using high molecular weight filters (100 kDa) for both dialysis tubing and centrifuge filters used for buffer exchange. The most successful attempt is shown in Figure 6.18, the monomeric protein (primary series) peaks are resolved, with the dimer partially and trimer and tetramer peaks being poorly separated. There are additional high charge peaks of monomeric mass; this is due to the metastability of the protein, as a secondary more open protein conformation would expose more basic side chains, leading to an increase in protonation.

These results show that it is unlikely that it will be possible to observe small molecule binding to the *ex vivo* oligomers, but this is not necessary for this experiment to be a success. Potential drug molecules for blocking and reversing aggregation, like Ac-TTAI-NH_2, could be added to aliquots of the sample and the relative abundance of signal in oligomer regions against the monomer signal could be compared. If the ligand is reversing aggregation then the oligomer abundances will drop indicating that the ligand is working as expected.

6.3.4 Ion mobility analysis of *ex vivo* polymers

Ekeowa *et al.* used ion mobility mass spectrometry to analyse the CCS percentage difference between monomeric and dimeric α_1-antitrypsin. They analysed heated M α_1-antitrypsin monomers and dimers as well as the loop-sheet and β hairpin models for polymerisation [37]. We have replicated this study as closely as possible, whilst including the new C-terminal model for polymerisation and using experimental CCS values from *ex vivo* Z mutant α_1-antitrypsin instead of data from the M variant.

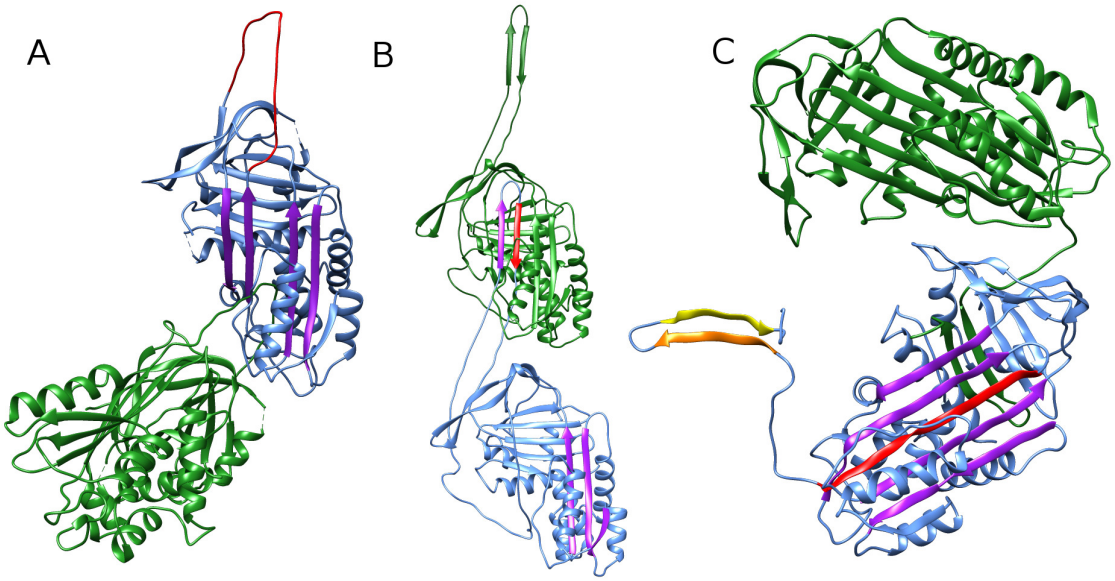

Figure 6.19: α_1-antitrypsin polymer models used to compare to *ex vivo* Z mutant dimeric α_1-antitrypsin. β-sheet A strands are coloured purple and the RCL is coloured red. The models shown are loop-sheet (A), β hairpin (B) and C-terminal domain swap (C). In model C the s4B strand is shown in orange and s5B in yellow.

The α_1-antitrypsin loop-sheet and β hairpin models used in the previous analysis [37] were kindly donated by Dr. Bibek Gooptu (Figure 6.19A and B) and the percentage difference in CCS was calculated from a monomeric crystal structure of α_1-antitrypsin. For this analysis a model was created for the C-terminal domain linkage, the crystal structure which was uploaded to the PDB was a single monomer out of the total trimer as the monomer was the repeating unit (PDB 3T1P). The SymmDock server [54] was used to get the structure of the trimer, and the PDB file was edited manually to remove one of the chains to create a dimer (Figure 6.19C).

The original study carried out the comparison with differing sequence coverages, with the loop-sheet missing N46 as well as having one residue missing from the N-terminus (N24) in comparison to the hairpin dimer. The models for hairpin dimer and C-terminal trimer start at N24 and finish at K394 with no missing residues. To maximise similarity with the original study it was decided to follow the hairpin dimer coverage, and so the loop-sheet and C-terminal domain swap model were used with amino acids N24-K394. The monomer structure (QLP1) used in the previous study had sequence coverage from F23-K394, and was also kept the same. The monomer pdf file contained solvent molecules, which were stripped before analysis. None of the other model files contained solvent.

The CCS values were calculated using Impact*, which is an implementation of the projection approximation algorithm [55]. The percentage increase in CCS (p) of dimer (Ω_d) from monomer (Ω_m), was calculated as shown in Equation 6.2, and the values were used to compare the differences found in models with experimental data.

$$p = \frac{\Omega_d}{\Omega_m} \cdot 100 \tag{6.2}$$

The peaks in the *ex vivo* Z α_1-antitrypsin mass spectrum are poorly separated (Figure 6.20A) and so thorough analysis of the m/z window used for ion mobility was required. First, the mass spectrum was collected (Figure 6.20A), and the peak region for each charge state was determined. Following this the sample analysed with quadrupole isolation, with iterative improvements being made before settling on the mass spectra shown in Figure 6.20B-G. Following this Amphitrite was used to select only the regions determined to be relevant to a particular charge state, with the final limits shown in Figure 6.20 as shaded regions.

The arrival time distributions extracted were converted to collision cross section distributions (CCSD), using Amphitrite, and the results are shown in Figure 6.21A. The CCS values of the different charge states are similar for the monomer and dimer distributions, exact values for the peak top CCS and

*http://impact.chem.ox.ac.uk/

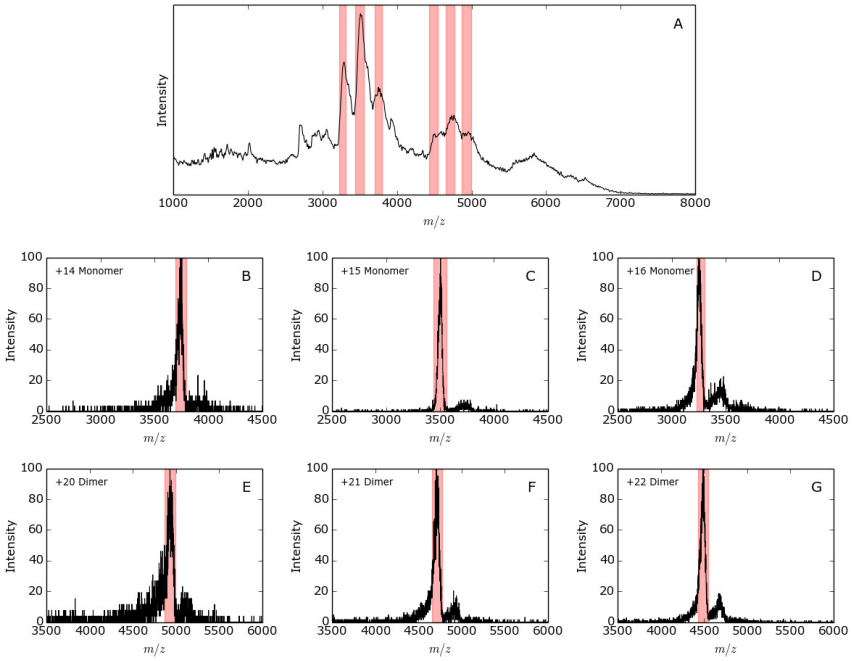

Figure 6.20: Mass spectra of *ex vivo* polymers of Z mutant α_1-antitrypsin, with data extraction limits used in ion mobility analysis. Full mass spectrum used for deciding quadrupole isolation settings (A) and ion mobility mass spectra of each charge state used for analysis (B-G).

weighted average are shown in Appendix 6.6.

The lowest charge state of the monomeric and dimeric protein were used to calculate the percentage increase in CCS between the two states and the result is shown as a red line in Figure 6.21 (for table of numerical values see Appendix 6.7). The experimental value is similar to that of the loop-sheet and C-terminal domain swap models of polymerisation, with the β hairpin being substantially larger. The percentage difference in CCS found by Ekeowa *et al.* (176 %) is similar to that of the Z (169 %) further supporting that they have the same structure.

The models used by Ekeowa *et al.* do however look unrealistic especially as gas-phase structures and so the C-terminal model may not be the most closely agreeing model. In the case of the loop-sheet model, the RCL of the blue molecule in Figure 6.19A is more extended than is likely to occur in the gas-phase where it would be likely that the loop would collapse to be closer to the molecule [56]. Additionally the RCL of the green protein is only partially

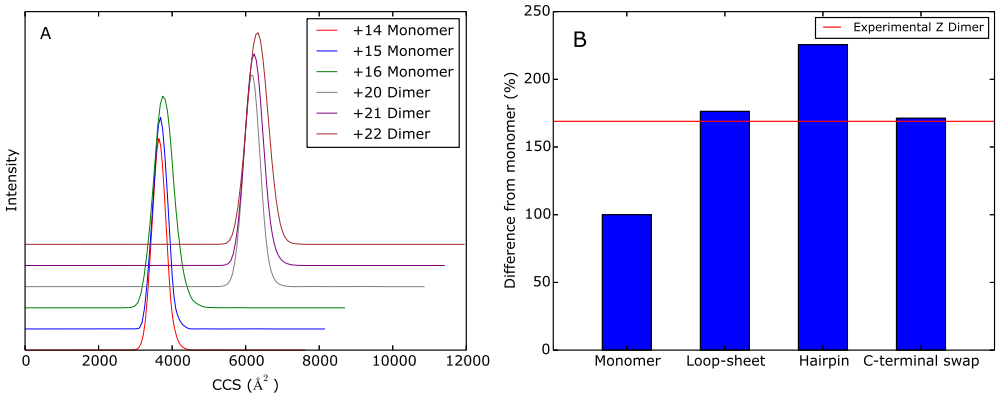

Figure 6.21: (A) CCS distributions for monomeric and dimeric *ex vivo* Z mutant α_1-antitrypsin liver polymers. (B) Bar chart showing the percentage difference in CCS from monomeric protein, PDB structure 1QLP is used for calculating values for the models (bars) and experimental Z mutant monomer CCS is used for calculating the percentage change to experimental dimer (red line).

inserted into the β-sheet. As the interaction is so strong it is likely that the RCL would insert more fully into the β-sheet, as seen in the protease binary complex structure, thereby pulling the two proteins closer together. These two features together would result in a lower CCS value being calculated for the dimer and so making the result closer to what was found experimentally. A similar situation can be seen with the β hairpin dimer, with the hairpin linkage of the green protein (Figure 6.19) fully extending away from the protein in a way that is highly unlikely in the gas-phase.

In order to improve on this work, our collaborators who specialise in molecular modelling will be creating more realistic structures for these molecules, using techniques such as structural energy minimisation, which can be compared with the experimental data. They will additionally be adding glycosylates to the models so that CCS values can be used in comparisons instead of percentage change in CCS. This could be of special importance if the monomeric Z protein has a higher observed CCS value as it is more unstable and could occupy a more unfolded monomeric conformation.

6.4 Conclusion

The work presented here has elucidated several characteristics of the interaction between α_1-antitrypsin and Ac-TTAI-NH_2, including that the peptide binds to the wild type form of the protein. The protein is now known to bind with a 2:1 peptide to protein ratio, the binding site is likely to be in the centre of β sheet A as the cleft has been shown to be large enough to bind two Ac-TTAI-NH_2 molecules and is logical as the peptide was originally designed to be an analog of the RCL. This interaction would mimic the standard interaction with a protease, which increases the stability of the protein. The unfolding studies have shown that the double bound form of α_1-antitrypsin has increased stability and reduced conformational variability, further supporting the conclusion. Furthermore the deconvolution of unfolding data suggests a specific pattern in the way in which Ac-TTAI-NH_2 stabilises α_1-antitrypsin. At low collision energies the apo form of the protein exists in two conformations whereas the holo form predominantly occupies a single conformation, the inference is then that the binding inhibits the occupation of a secondary conformation, which is potentially related to the aggregation process.

The interaction has been found to be cooperative, this is likely due to the first binding interaction opening the cleft in β sheet A enabling the second peptide to bind more readily. This explains why in the initial experiments, only a single binding event was detected as the second binding event would occur rapidly after the first, leading the researchers to believe that in the case of the Z mutant a single peptide binds, which now seems unlikely.

In developing a framework for future drug discovery experiments, this chapter introduces work with Z mutant protein. As it is not possible to express monomeric Z mutant α_1-antitrypsin in *E. coli*, the protein has been obtained from human samples and hence is glycosylated. Using mass spectral simulation it has been shown that, with additional sample, it would be possible to carry out ion mobility unfolding experiments to analyse the interaction Ac-TTAI-NH_2 and Z mutant α_1-antitrypsin.

Additionally analyses of glycosylated α_1-antitrypsin oligomers were con-

ducted. It was shown that using monomeric wild type protein extracted from human plasma, which was polymerised *in vitro*, it was possible to get peak separation for oligomeric states up to tetramers. The success of this analysis means that it should at least be possible to monitor the reversal of polymerisation in the presence of drug molecules if not to actually get adequate peak separation to carry out quadrupole isolation experiments.

Finally the first mass spectrum of an *ex vivo* polymer is presented. The sample was extracted from hepatocyte inclusion bodies, and the sample contains high concentrations of impurities. It is, however, possible to determine the *m/z* regions belonging to monomer, dimer, trimer and tetramer. After having screened M, wild type and/or plasma Z with potential drug molecules for binding using Amphitrite, and assessed increases in stability using Challenger; it would be possible to test the best molecules on a sample like this one. The relative abundance of monomer to oligomer could be analysed and this would give strong evidence as to whether the drug molecule could reverse polymerisation and clear out the inclusion bodies of patients with severe α_1-antitrypsin deficiency syndrome.

The ion mobility mass spectrometry analysis of *ex vivo* Z polymers was successful and this will hopefully lead to the acquisition of more protein which will allow replication of the results, thereby increasing accuracy. It was shown that the percentage change in CCS was similar for Z *ex vivo* dimer and dimeric M from heating, supporting evidence that the two proteins have similar dimeric structure. More accurate readings could then be compared to ion mobility CCS values of the M polymer further supporting this conclusion in terms of absolute CCS, thereby removing the potential factor of variation in monomer structure.

It was found that the C-terminal model for polymerisation had the most similar change in CCS value to those experimentally derived. Hopefully the results will lead to further work into the modelling aspect of the analysis allowing for a more robust comparison between the models and experimental data.

Contributions

All mass spectrometry experiments and data analysis was conducted by Ganesh N. Sivalingam. Z mutant α_1-antitrypsin as purified from plasma was provided by Dr. Imran Haq. Extraction of *ex vivo* Z polymers from liver samples, and M α_1-antitrypsin from plasma, was carried out by Sarah Faull. Preparing samples for mass spectrometry experiments was conjointly conducted by Sarah Faull and Ganesh Sivalingam.

References

[1] Nyon, M. P., Segu, L., Cabrita, L. D., Lévy, G. R., Kirkpatrick, J., Roussel, B. D., Patschull, A. O., Barrett, T. E., Ekeowa, U. I., Kerr, R., Waudby, C. A., Kalsheker, N., Thalassinos, K., Lomas, D. A., Christodoulou, J., and Gooptu, B. (2012). "Structural dynamics associated with intermediate formation in an archetypal conformational disease". *Structure* 20.3, pp. 504–512.

[2] Gooptu, B. and Lomas, D. A. (2009). "Conformational Pathology of the Serpins: Themes, Variations, and Therapeutic Strategies". *Annual Review of Biochemistry* 78.1, pp. 147–176.

[3] Huntington, J. A., Read, R. J., and Carrell, R. W. (2000). "Structure of a serpin–protease complex shows inhibition by deformation". *Nature* 407.6806, pp. 923–926.

[4] Lomas, D. A. and Carrell, R. W. (2002). "Serpinopathies and the conformational dementias". *Nature Reviews Genetics* 3.10, pp. 759–768.

[5] Brantly, M., Nukiwa, T., and Crystal, R. G. (1988). "Molecular basis of alpha-1-antitrypsin deficiency". *The American Journal of Medicine* 84.6, Supplement 1, pp. 13–31.

[6] Parfrey, H., Mahadeva, R., and Lomas, D. A. (2003). "α_1-antitrypsin deficiency, liver disease and emphysema". *The International Journal of Biochemistry & Cell Biology* 35.7, pp. 1009–1014.

[7] Jacobsson, K. (1955). "I. Studies on the determination of fibrinogen in human blood plasma. II. Studies on the trypsin and plasmin inhibitors in human blood serum". *Scandinavian Journal of Clinical and Laboratory Investigation* 7 Suppl. 14, pp. 3–102.

[8] Laurell, C.-B. and Eriksson, S. (1963). "The electrophoretic α_1-globulin pattern of serum in α_1-antitrypsin deficiency". *Scandinavian Journal of Clinical & Laboratory Investigation* 15.2, pp. 132–140.

[9] Sharp, H., Bridges, R., Krivit, W., and Freier, E. (1969). "Cirrhosis associated with alpha-1-antitrypsin deficiency: a previously unrecognized inherited disorder." *The Journal of laboratory and clinical medicine* 73.6, pp. 934–939.

[10] Johnson, A. M. and Alper, C. A. (1970). "Deficiency of α_1-antitrypsin in childhood liver disease". *Pediatrics* 46.6, pp. 921–925.

[11] Chua, F. and Laurent, G. J. (2006). "Neutrophil Elastase". *Proceedings of the American Thoracic Society* 3.5, pp. 424–427.

[12] Larsson, C. (1978). "Natural history and life expectancy in severe alpha$_1$-antitrypsin deficiency, Pi Z". *Acta Medica Scandinavica* 204.1-6, pp. 345–351.

[13] Massi, G and Chiarelli, C (1994). "Alpha$_1$-antitrypsin: molecular structure and the Pi system". *Acta Pædiatrica* 83, pp. 1–4.

[14] Blanco, I, Fernández, E, and Bustillo, E. (2001). "Alpha-1-antitrypsin PI phenotypes S and Z in Europe: an analysis of the published surveys". *Clinical Genetics* 60.1, pp. 31–41.

[15] Hutchison, D. C. S., Tobin, M. J., and Cook, P. J. L. (1983). "Alpha1-antitrypsin deficiency: Clinical and physiological features in heterozygotes of Pi types SZ: A survey by the British Thoracic Association". *British Journal of Diseases of the Chest* 77, pp. 28–34.

[16] Sinden, N. J., Koura, F., and Stockley, R. A. (2014). "The significance of the F variant of alpha-1-antitrypsin and unique case report of a PiFF homozygote". *BMC Pulmonary Medicine* 14.1, p. 132.

[17] Stoller, J. K. and Aboussouan, L. S. (2005). "α_1-antitrypsin deficiency". *The Lancet* 365.9478, pp. 2225–2236.

[18] Ogushi, F., Hubbard, R. C., Fells, G. A., Casolaro, M. A., Curiel, D. T., Brantly, M. L., and Crystal, R. G. (1988). "Evaluation of the S-type of alpha-1-antitrypsin as an *in vivo* and *in vitro* inhibitor of neutrophil elastase". *American Review of Respiratory Disease* 137.2, pp. 364–370.

[19] Kemmer, N., Kaiser, T., Zacharias, V., and Neff, G. W. (2008). "Alpha-1-Antitrypsin Deficiency: Outcomes After Liver Transplantation". *Transplantation Proceedings* 40.5, pp. 1492–1494.

[20] Dawkins, P. A., Dowson, L. J., Guest, P. J., and Stockley, R. A. (2003). "Predictors of mortality in α_1-antitrypsin deficiency". *Thorax* 58.12, pp. 1020–1026.

[21] Stoller, J. K. (2002). "Acute Exacerbations of Chronic Obstructive Pulmonary Disease". *New England Journal of Medicine* 346.13, pp. 988–994.

[22] Sutherland, E. R. and Cherniack, R. M. (2004). "Management of Chronic Obstructive Pulmonary Disease". *New England Journal of Medicine* 350.26, pp. 2689–2697.

[23] Gadek, J. E., Klein, H. G., Holland, P. V., and Crystal, R. G. (1981). "Replacement therapy of alpha 1-antitrypsin deficiency. Reversal of protease-antiprotease imbalance within the alveolar structures of PiZ subjects." *Journal of Clinical Investigation* 68.5, pp. 1158–1165.

[24] Yu, M.-H., Lee, K. N., and Kim, J. (1995). "The Z type variation of human α_1-antitrypsin causes a protein folding defect". *Nature Structural & Molecular Biology* 2.5, pp. 363–367.

[25] Dickens, J. A. and Lomas, D. A. (2011). "Why has it been so difficult to prove the efficacy of alpha-1-antitrypsin replacement therapy? Insights from the study of disease pathogenesis". *Drug Design, Development and Therapy* 5, pp. 391–405.

[26] Wewers, M. D., Casolaro, M. A., Sellers, S. E., Swayze, S. C., McPhaul, K. M., Wittes, J. T., and Crystal, R. G. (1987). "Replacement therapy for alpha1-antitrypsin deficiency associated with emphysema". *New England Journal of Medicine* 316.17, pp. 1055–1062.

[27] Stoller, J. K., Snider, G. L., Brantly, M. L., Fallat, R. J., Stockley, R. A., Turino, G. M., Konietzko, N., Dirksen, A., Eden, E., Fallat, R. J., Luisetti, M., Stolk, J., and Strange, C. (2005). "American Thoracic So-

ciety/European Respiratory Society Statement: Standards for the Diagnosis and Management of Individuals with Alpha-1 Antitrypsin Deficiency". *Pneumologie* 59.1, pp. 36–68.

[28] Lomas, D. A. (2006). "The selective advantage of α_1-antitrypsin deficiency". *American Journal of Respiratory and Critical Care Medicine* 173.10, pp. 1072–1077.

[29] Cichy, J., Potempa, J., and Travis, J. (1997). "Biosynthesis of α_1-proteinase inhibitor by human lung-derived epithelial cells". *Journal of Biological Chemistry* 272.13, pp. 8250–8255.

[30] Perlmutter, D. H., Daniels, J. D., Auerbach, H. S., Schryver-Kecskemeti, K. D., Winter, H. S., and Alpers, D. H. (1989). "The alpha 1-antitrypsin gene is expressed in a human intestinal epithelial cell line." *Journal of Biological Chemistry* 264.16, pp. 9485–9490.

[31] Dafforn, T. R., Mahadeva, R., Elliott, P. R., Sivasothy, P., and Lomas, D. A. (1999). "A kinetic mechanism for the polymerization of α_1-antitrypsin". *Journal of Biological Chemistry* 274.14, pp. 9548–9555.

[32] Elliott, P. R., Bilton, D., and Lomas, D. A. (1998). "Lung Polymers in Z α_1-Antitrypsin Deficiency-related Emphysema". *American Journal of Respiratory Cell and Molecular Biology* 18.5, pp. 670–674.

[33] Mulgrew, A. T., Taggart, C. C., Lawless, M. W., Greene, C. M., Brantly, M. L., O'Neill, S. J., and McElvaney, N. G. (2004). "Z α_1-antitrypsin polymerizes in the lung and acts as a neutrophil chemoattractant". *Chest* 125.5, pp. 1952–1957.

[34] Parmar, J. S., Mahadeva, R., Reed, B. J., Farahi, N., Cadwallader, K. A., Keogan, M. T., Bilton, D., Chilvers, E. R., and Lomas, D. A. (2002). "Polymers of α_1-antitrypsin are chemotactic for human neutrophils". *American Journal of Respiratory Cell and Molecular Biology* 26.6, pp. 723–730.

[35] Proctor, R. N. (2004). "The Global Smoking Epidemic: A History and Status Report". *Clinical Lung Cancer* 5.6, pp. 371–376.

[36] Abouharb, M. R. and Kimball, A. L. (2007). "A New Dataset on Infant Mortality Rates, 1816—2002". *Journal of Peace Research* 44.6, pp. 743–754.

[37] Ekeowa, U. I., Freeke, J., Miranda, E., Gooptu, B., Bush, M. F., Pérez, J., Teckman, J., Robinson, C. V., and Lomas, D. A. (2010). "Defining the mechanism of polymerization in the serpinopathies". *Proceedings of the National Academy of Sciences* 107.40, pp. 17146–17151.

[38] Lomas, D. A., LI-Evans, D., Finch, J. T., and Carrell, R. W. (1992). "The mechanism of Z α_1-antitrypsin accumulation in the liver". *Nature* 357.6379, pp. 605–607.

[39] James, E. L. and Bottomley, S. P. (1998). "The mechanism of α_1-antitrypsin polymerization probed by fluorescence spectroscopy". *Archives of Biochemistry and Biophysics* 356.2, pp. 296–300.

[40] Gooptu, B., Hazes, B., Chang, W.-S. W., Dafforn, T. R., Carrell, R. W., Read, R. J., and Lomas, D. A. (2000). "Inactive conformation of the serpin α_1-antichymotrypsin indicates two-stage insertion of the reactive loop: Implications for inhibitory function and conformational disease". *Proceedings of the National Academy of Sciences* 97.1, pp. 67–72.

[41] Yamasaki, M., Li, W., Johnson, D. J. D., and Huntington, J. A. (2008). "Crystal structure of a stable dimer reveals the molecular basis of serpin polymerization". *Nature* 455.7217, pp. 1255–1258.

[42] Miranda, E., Pérez, J., Ekeowa, U. I., Hadzic, N., Kalsheker, N., Gooptu, B., Portmann, B., Belorgey, D., Hill, M., Chambers, S., Teckman, J., Alexander, G. J., Marciniak, S. J., and Lomas, D. A. (2010). "A novel monoclonal antibody to characterize pathogenic polymers in liver disease associated with α_1-antitrypsin deficiency". *Hepatology* 52.3, pp. 1078–1088.

[43] Yamasaki, M., Sendall, T. J., Pearce, M. C., Whisstock, J. C., and Huntington, J. A. (2011). "Molecular basis of α_1-antitrypsin deficiency revealed by the structure of a domain-swapped trimer". *EMBO Reports* 12.10, pp. 1011–1017.

[44] Chang, Y.-P. and Chu, Y.-H. (2013). "Blocking formation of large protein aggregates by small peptides". *Chemical Communications* 49.41, pp. 4591–4600.

[45] Mallya, M., Phillips, R. L., Saldanha, S. A., Gooptu, B., Leigh Brown, S. C., Termine, D. J., Shirvani, A. M., Wu, Y., Sifers, R. N., Abagyan, R., and Lomas, D. A. (2007). "Small Molecules Block the Polymerization of Z α_1-Antitrypsin and Increase the Clearance of Intracellular Aggregates". *Journal of Medicinal Chemistry* 50.22, pp. 5357–5363.

[46] Schulze, A. J., Baumann, U., Knof, S., Jaeger, E., Huber, R., and Laurell, C.-B. (1990). "Structural transition of α_1-antitrypsin by a peptide sequentially similar to β-strand s4A". *European Journal of Biochemistry* 194.1, pp. 51–56.

[47] Mahadeva, R., Dafforn, T. R., Carrell, R. W., and Lomas, D. A. (2002). "6-mer peptide selectively anneals to a pathogenic serpin conformation and blocks polymerization; Implications for the prevention of z α_1-antitrypsin-related cirrhosis". *Journal of Biological Chemistry* 277.9, pp. 6771–6774.

[48] Chang, Y.-P., Mahadeva, R., Chang, W.-S. W., Shukla, A., Dafforn, T. R., and Chu, Y.-H. (2006). "Identification of a 4-mer peptide inhibitor that effectively blocks the polymerization of pathogenic z α_1-antitrypsin". *American Journal of Respiratory Cell and Molecular Biology* 35.5, pp. 540–548.

[49] Chang, Y.-P., Mahadeva, R., Chang, W.-S. W., Lin, S.-C., and Chu, Y.-H. (2009). "Small-molecule peptides inhibit Z α_1-antitrypsin polymerization". *Journal of Cellular and Molecular Medicine* 13.8b, pp. 2304–2316.

[50] Bolmer, S. and Kleinerman, J. (1986). "Isolation and characterization of α_1-antitrypsin in PAS-positive hepatic granules from rats with experimental α_1-antitrypsin deficiency." *The American Journal of Pathology* 123.2, pp. 377–389.

[51] Eriksson, S. and Larsson, C. (1975). "Purification and partial characterization of pas-positive inclusion bodies from the liver in alpha1-antitrypsin deficiency". *New England Journal of Medicine* 292.4, pp. 176–180.

[52] Steward, M. W. and Petty, R. E. (1972). "The use of ammonium sulphate globulin precipitation for determination of affinity of anti-protein antibodies in mouse serum". *Immunology* 22.5, pp. 747–756.

[53] Edwards, M. J., Williams, M. A., Maxwell, A., and McKay, A. R. (2011). "Mass Spectrometry Reveals That the Antibiotic Simocyclinone D8 Binds to DNA Gyrase in a "Bent-Over" Conformation: Evidence of Positive Cooperativity in Binding". *Biochemistry* 50.17, pp. 3432–3440.

[54] Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. (2005). "PatchDock and SymmDock: servers for rigid and symmetric docking". *Nucleic Acids Research* 33.suppl 2, W363–W367.

[55] Mesleh, M. F., Hunter, J. M., Shvartsburg, A. A., Schatz, G. C., and Jarrold, M. F. (1996). "Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential". *The Journal of Physical Chemistry* 100.40, pp. 16082–16086.

[56] Jurneczko, E. and Barran, P. E. (2010). "How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase". *Analyst* 136.1, pp. 20–28.

6.5 Appendix

6.5.1 Analysis of dissociated holo in CIU experiments

The dissociation products of the CIU experiment are analysed here, and the results are shown in Figure 6.22. As expected the mass spectra of the apo protein changes very little as the collision energy increases, with no reduced charge species appearing (Figure 6.22A).

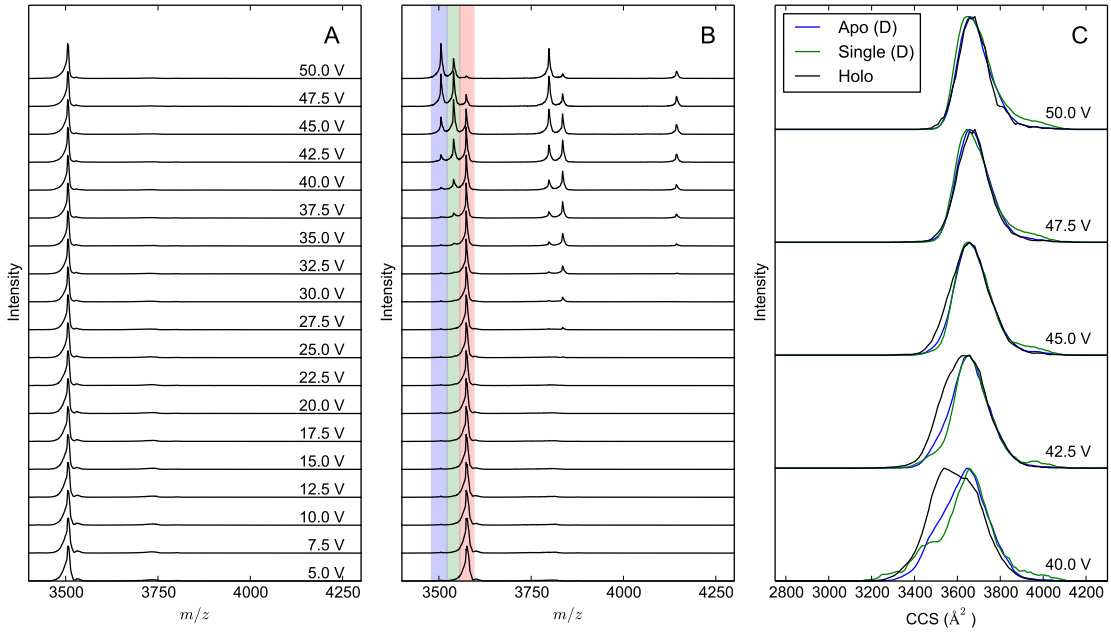

Figure 6.22: Analysis of dissociated holo in the CIU experiment. The mass spectra of the apo protein throughout the CIU experiment are shown in A. The mass spectra for the holo protein are shown in B, with the shaded regions showing the m/z regions for the +13 apo (blue), single bound (green) and holo (black). The m/z regions highlighted in B are extracted to show the CCS distributions for the dissociated apo and single bound protein, and are displayed in C.

The mass spectra of the holo experiment are shown in Figure 6.22B. The holo protein is stable until 27.5 V, where dissociation begins to occur to a detectable level. Between 27.5 and 35 V, the dissociation primarily occurs via the dissociation of one Ac-TTAI-NH_2 molecule with a proton, resulting in a peak for single bound +12 α_1-antitrypsin. As the voltages increase there also appears a +12 apo peak, meaning that both Ac-TTAI-NH_2 molecules have dissociated, whilst removing only one charge. At higher voltages a peak

corresponding to +11 apo α_1-antitrypsin also appears, indicating that two Ac-TTAI-NH_2 molecules dissociated with 2 charges. Interestingly no single bound protein is found with a charge state of +11, which means that Ac-TTAI-NH_2 does not dissociate with two protons, to a detectable level.

Examining the +13 apo and single bound species shows that as the voltage increases, the single bound species is initially more abundant than the double bound, with the double bound species becoming most abundant at the highest collision energies.

When the collision energy is above 37.5 V, there is sufficient signal for ion mobility analysis of the +13 apo and single bound species, and the results are shown in Figure 6.22C. For 47.5 and 50 V collision energies, there is a very close similarity in the CCSDs of all three bound states, with single bound seeming to be starting to occupy an even more unfolded conformation.

The CCSDs corresponding to 40 and 42.5 V show differences between the species. At 40 V the apo and single bound forms are close in CCS to the final unfolded holo form of the protein. This is of interest as it is likely that the conformation that allowed for the initial dissociation was the most unfolded conformation.

The apo and holo forms of the protein do not occupy any more unfolded conformational families than the most abundant seen at 50 V. The single bound, however, appears to be unfolding further, with an additional conformation being detectable at 3955 $Å^2$. This suggests that the single bound form is more unstable than the apo form, and could explain why the binding interaction is cooperative.

6.5.2 Ac-TTAI-NH_2 - α_1-antitrypsin titration

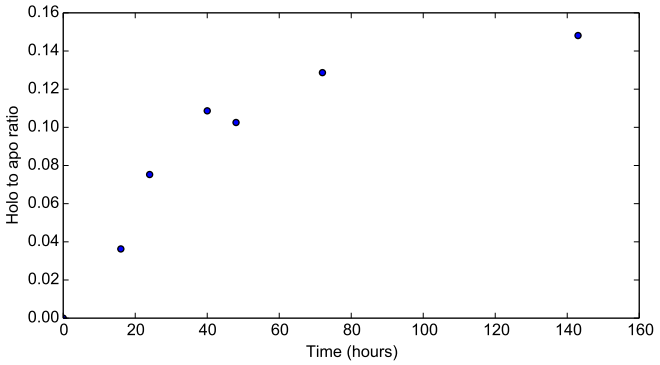

Figure 6.23: Time course of wild type α_1-antitrypsin binding to Ac-TTAI-NH_2. The experiment compares apo with double bound intensities at a peptide:protein ratio of 0.1:1.

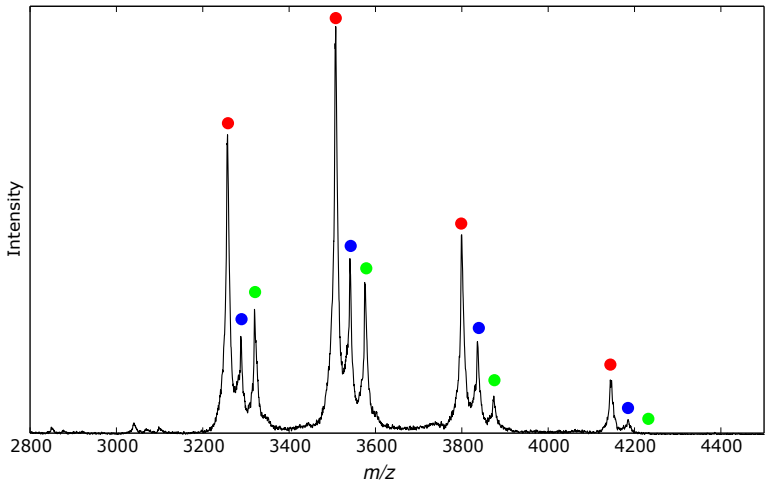

Figure 6.24: Mass spectrum of Ac-TTAI-NH_2 bound to wild type α_1-antitrypsin after incubation and buffer exchange. Peaks marked with red dots indicate apo protein, with single and double bound α_1-antitrypsin respectively marked with blue and green dots.

6.5.3 Tables of CCS values for *ex vivo* analysis

Oligomer	Charge	Peak CCS (Å^2)	Mean CCS (Å^2)
Monomer	+14	3,639.1	3,669.6
Monomer	+15	3,684.6	3,700.7
Monomer	+16	3,744.3	3,821.9
Dimer	+20	6,148.8	6,191.0
Dimer	+21	6,229.6	6,262.6
Dimer	+22	6,339.1	6,345.8

Table 6.6: Table of collision cross section (CCS) values of monomeric and dimeric charge states of *ex vivo* polymer Z mutant α_1-antitrypsin, given as peak top values and weighted mean CCS values.

Protein	CCS (Å^2)	Monomer CCS (%)
Experimental monomer	3,639	100.0
Experimental dimer	6,149	169.0
1QLP Monomer	2,695	100.0
Loop-sheet	4,753	176.4
β-hairpin	6,078	225.5
C-terminal swap	4,617	171.3

Table 6.7: Table of the CCS values and percentage increase in CCS for three polymerisation models and experimental data of *ex vivo* Z mutant α_1-antitrypsin polymer.

Chapter 7

Conclusions

7.1 Amphitrite

Amphitrite introduces a new method for deconvoluting mass spectra, which includes a graphical interface for ease of use. This functionality was important throughout the work presented here. A particularly important example was the analysis of α_1-antitrypsin-Ac-TTAI-NH_2 binding where it offered the benefit of greater accuracy for bound state abundance analysis.

Several new methods for IM-MS data analysis were also introduced; including spectral averaging for creating representative spectra for replicated data and heatmap difference plots for the comparison of analytes or analyte conditions. The automated calibration functionality was the most useful aspect for the work presented in this thesis. In addition to automating the extraction of calibrant data and calculation of calibration curves, it allowed for the automatic conversion of ATDs to CCSDs, and the ability to convert the arrival time axis on heatmaps to CCS.

The work for Amphitrite, early in my Ph.D., facilitated rapid data analysis and the progress made herein, was as a result of this development process. There are, however, still areas in which the project could be further improved. The back-end of the software is very strong, as it has seen continual use and improvement producing statistically robust and accurate data; however, the

graphical user interface is still at a prototype stage. It was too large a project to be perfected by a single person within the available time, especially with obligations to complete several other projects. It is my hope that the work of Amphitrite will inspire additional work in the area. As the software is open source, the back-end could be built upon directly, requiring only an improved interface to ensure ease and accessibility to the tecnical aspects of the tools within for a wider range of researchers.

The mass spectral deconvolution aspect has worked well, but for highly complex mass spectrum, the user input can still be quite time consuming. It seems possible to accomplish a fully automated deconvolution, and there is more than one way this could be implemented. Tseng *et al.* introduced a method of automatic mass assignment [1], this method could be used as initial information for a deconvolution algorithm. A second method could be to utilise the arrival time data of TWIM-MS data. Figure 3.4 shows the characteristic non-linear curve in travelling wave arrival time for the charge states of a particular ion. A two-dimensional deconvolution method could use the arrival time information to group together the charge states of a peak and the mass spectral deconvolution to assign mass and peak characteristics. The work contained herein is but a step from which a diversity of improvements could proceed.

7.2 Challenger

The open source Amphitrite back-end allows for the programmatic manipulation of TWIM-MS data files, and this was used to develop the Challenger algorithm and associated tools. New methods were introduced that summarise ATDs allowing for quantitative assessment of the degree of unfolding for use with gas-phase unfolding data. The methods were applied to the unfolding of apo α_1-antitrypsin and when bound to Ac-TTAI-NH_2. The results obtained showed clear differences in the gas-phase stability and conformational variability between the bound states, confirming the hypothesis that Ac-TTAI-NH_2 blocks α_1-antitrypsin polymerisation by stabilising the protein.

The Challenger algorithm is a purpose built genetic algorithm for the deconvolution of gas-phase unfolding data. It calculates the centre, and so CCS, of a conformation more accurately than would be determined by eye. It uses the position of a conformation in the multiple ATDs produced by unfolding experiments to accurately assign the centre. The deconvolution methodology allows for a much more accurate determination of the abundance of different conformations in comparison to peak height analysis. This is used to track the proportional abundances of conformations throughout unfolding experiments. When this methodology was applied to α_1-antitrypsin and Ac-TTAI-NH_2, it revealed clear distinctions in the pattern of unfolding between apo and holo α_1-antitrypsin that were not detectable using CIU fingerprints. This information can be used when selecting new ligands; if the pattern of unfolding is similar to holo α_1-antitrypsin, it is likely that the interaction stabilises the protein with a similar binding mechanism to Ac-TTAI-NH_2, and so would be likely to block α_1-antitrypsin polymerisation.

The Challenger chapter also introduces a method for automating the acquisition of IM-MS unfolding experiments. These experiments can be very challenging, as it is preferable that the data for all analytes are collected in the same experimental session as well as the data for the IM calibration. The automation means that operators can have time to think and plan while the unfolding experiment is running, and I found personally that this substantially increased the success rate of often tumultuous experimental sessions.

In the future, the Challenger algorithm could be adapted to a multi-stage deconvolution. A statistic could be created that would group together sets of similar ATDs, creating an initial reduced dataset. Spectral averaging could then be used to create a representative ATD from the members of the group. This reduced dataset could then be analysed using the current Challenger algorithm, greatly reducing the run time and ameliorating issues caused by overrepresentation of ATD features during conformational centre determination. This could then be followed by a second algorithm, which does not optimise for the conformational centre. This would then be applied to the entire dataset using the conformational centres determined in the previous step, once again with a reduced run time in comparison to the current Challenger

implementation.

7.3 α_1-antitrypsin

α_1-antitrypsin can form polymers due to deleterious point mutations, and the polymerisation can be blocked by Ac-TTAI-NH_2 binding. Presented here is evidence that demonstrates that the peptide binds with a 2:1 stoichiometry, raising questions about the existing binary complex model. The interaction has been analysed by a variety of IM-MS techniques in conjunction with the new data analysis techniques presented in this thesis. The results show that IM-MS can be a viable technique for high-throughput drug screening towards finding alternative small molecules with the ability to block α_1-antitrypsin polymerisation.

The first IM-MS analysis of a protein aggregate extracted from a patient is also presented. This was in an effort to further understand the mechanism of polymerisation, a highly contested question in the α_1-antitrypsin field. It was found that the loop-sheet and C-terminal domain swap models best fit the IM-MS data, and it may be the case that polymers formed by both mechanisms are present in a disease state liver. Future work will include improvements to the coordinate files used to represent the models, preferably with the inclusion of glycosylates, for a better comparison to the samples used in the IM-MS analysis. With access to more of this sample, it would be possible to use *ex vivo* polymers as the final stage of mass spectrometry testing after drug screening analysis.

7.4 Final remarks

Software development for the analysis of IM-MS data is starting to become more abundant and it seems like this could be an important area for the expansion of the field. This movement will bring new challenges that I have already experienced during my Ph.D. The most important development, in my opinion, would be an open data format with easy to use and freely available conversion tools. The mass spectrometry proteomics field has mzMl for this purpose [2], and there are several tools available for converting data from different mass spectrometer manufacturers to the same format. These data are in a format that is based on XML(eXtensible Markup Language), which is plain text and so can be easily accessed and manipulated. This is in stark contrast to the binary files used by instruments such as the Waters Synapt, which cannot be read without proprietary libraries. The most difficult challenge with the development of Amphitrite has been the access to the Synapt data. The current iteration of our software works with data from the Synapt G1, but not with newer instruments. Additionally, we are under the impression that we are not permitted to distribute the MS Windows-only proprietary library that is required to open the data files. This means that users have to find the library file and copy it to a specific location. The problem is further compounded by the fact that it only works with the version of the library which ships with the Synapt G1 version of Driftscope. Working together with Waters (and potentially other manufacturers), this problem could be overcome and would greatly reduce the barrier for entry for non-computational mass spectrometrists to employ the full functionality of software like Amphitrite.

Publications featuring an X-ray crystallography structure are now required to deposit the coordinate file to the publically available Protein Databank. This has been very successful, and there are many researchers who work directly with the PDB, without ever having carried out an X-ray crystallography experiment. It allows many levels of meta-analysis, sometimes involving thousands of coordinate files. This would never have been possible if the same laboratory doing the analysis also had to create all of the crystal structures.

The multidimensional data produced by IM-MS experiments contains a

wealth of information, and in my opinion would benefit greatly from a public data repository. Statisticians and computer scientists could benefit the field by creating new tools and methods of analysis, without having to purchase and learn to use the required equipment. Once again the, currently impossible, potential amount of data that could be analysed together could bring large advances in the understanding of mass spectral data. A salient example that comes to mind is that in 1989 Mann and Fenn said of ESI "the spectrum comprises a sequence of peaks with an intensity distribution that is near Gaussian" [3], while this is true, the distribution of peak heights is not truly Gaussian. A large publically available dataset of IM-MS data would make it straightforward for a mathematically minded scientist to solve this fundamental aspect of ESI mass spectra.

References

[1] Tseng, Y.-H., Uetrecht, C., Heck, A. J., and Peng, W.-P. (2011). "Interpreting the charge state assignment in electrospray mass spectra of bioparticles". *Analytical Chemistry* 83.6, pp. 1960–1968.

[2] Deutsch, E. (2008). "mzML: A single, unifying data format for mass spectrometer output". *Proteomics* 8.14, pp. 2776–2777.

[3] Mann, M., Meng, C. K., and Fenn, J. B. (1989). "Interpreting mass spectra of multiply charged ions". *Analytical Chemistry* 61.15, pp. 1702–1708.

Chapter 8

Publications

[1] Sivalingam, G. N., Yan, J., Sahota, H., and Thalassinos, K. (2013). "Amphitrite: A program for processing travelling wave ion mobility mass spectrometry data". *International Journal of Mass Spectrometry* 345–347, pp. 54–62.

[2] Wojnowska, M., Yan, J., Sivalingam, G. N., Cryar, A., Gor, J., Thalassinos, K., and Djordjevic, S. (2013). "Autophosphorylation Activity of a Soluble Hexameric Histidine Kinase Correlates with the Shift in Protein Conformational Equilibrium". *Chemistry & Biology* 20.11, pp. 1411–1420.

[3] Warelow, T. P., Oke, M., Schoepp-Cothenet, B., Dahl, J. U., Bruselat, N., Sivalingam, G. N., Leimkühler, S., Thalassinos, K., Kappler, U., Naismith, J. H., and Santini, J. M. (2013). "The respiratory arsenite oxidase: structure and the role of residues surrounding the Rieske cluster". *PloS One* 8.8, e72535.

[4] Sivalingam, G. N. and Shepherd, A. J. (2012). "An analysis of B-cell epitope discontinuity". *Molecular Immunology* 51.3–4, pp. 304–309.